

Билл
Фрэнкс

УКРОЩЕНИЕ БОЛЬШИХ ДАННЫХ

Это книга об анализе больших данных — о том, что они собой представляют и как их использовать в организации для повышения ее эффективности и конкурентоспособности.

*Томас Дэвенпорт,
сооснователь и директор по исследованиям
Международного института аналитики*

Как
извлекать
знания
из массивов
информации
с помощью
глубокой
аналитики

Эту книгу хорошо дополняют:

Большие данные

Виктор Майер-Шенбергер

Великий переход

Николас Карр

Новый цифровой мир

Эрик Шмидт

Bill Franks

Taming the Big Data Tidal Wave

Finding Opportunities in Huge Data
Streams with Advanced Analytics

John Wiley & Sons, Inc.

Билл Фрэнкс

Укрощение больших данных

Как извлекать знания
из массивов информации
с помощью глубокой аналитики

Перевод с английского Андрея Баранова

Издательство «Манн, Иванов и Фербер»
Москва, 2014

УДК 330.47
ББК 65.051.03
Ф93

Фрэнкс, Билл

Ф93 Укрощение больших данных: как извлекать знания из массивов информации с помощью глубокой аналитики / Билл Фрэнкс ; пер. с англ. Андрея Баранова. — М. : Манн, Иванов и Фербер, 2014. — 352 с.

ISBN 978-5-00057-146-0

По убеждению Билла Фрэнкса, ведущего аналитика всемирно известной компании Teradata, уже сейчас наступила эпоха совершенно новых подходов в аналитической сфере и в использовании больших объемов данных. Что такое большие данные, каково их значение, каковы методы, технологии и принципы новейшей аналитики и как это повлияет на последующее развитие бизнеса — в этой книге вы найдете подробную, четко структурированную, изложенную простым языком и наиболее полную информацию об этом явлении.

УДК 330.47
ББК 65.051.03

Все права защищены.
Никакая часть данной книги не может быть воспроизведена в какой бы то ни было форме без письменного разрешения владельцев авторских прав.
Правовую поддержку издательства обеспечивает юридическая фирма «Вегас-Лекс»

VEGAS LEX

© 2012 Bill Franks
© Перевод на русский язык, издание на русском языке, оформление. ООО «Манн, Иванов и Фербер», 2014

ISBN 978-5-00057-146-0

Оглавление

От партнера издания	9
Предисловие.....	13
Введение.....	19
ЧАСТЬ I. Появление больших данных	
Глава 1. Что такое «большие данные» и каково их значение?	29
Глава 2. Веб-данные: первые большие данные.....	55
Глава 3. Источники больших данных и их ценность	79
ЧАСТЬ II. Укрощение больших данных: технологии, процессы и методы	
Глава 4. Эволюция масштабируемости аналитических систем	115
Глава 5. Эволюция аналитических процессов.....	151
Глава 6. Эволюция аналитических инструментов и методов	185
ЧАСТЬ III. Укрощение больших данных: люди и подходы	
Глава 7. Что такое хороший анализ?	213
Глава 8. Что такое хороший профессионал в области аналитики?	235
Глава 9. Что такое хорошая аналитическая команда?	259
ЧАСТЬ IV. Объединение пройденного: аналитическая культура	
Глава 10. Создание условий для внедрения инноваций в сфере аналитики.....	283
Глава 11. Создание культуры инноваций и открытий	303
Заключение.....	325
Благодарности.....	329
Об авторе.....	331
Предметный указатель	333
Примечания	339

От партнера издания

Сегодня понятие «большие данные», бесспорно, очень популярно. Вокруг них создан огромный ажиотаж, многие действительно связывают с ними будущее. Но есть и те, кто настроен скептически или с осторожностью к большим данным.

Дело в том, что под этим модным выражением сегодня продают самое разное содержание. Одни считают это абсолютно новым революционным технологическим прорывом, подразумевающим полную замену существующих технологий и методологий. Другие — лишь логичным дополнением и развитием старого устойчивого тренда бизнес-аналитики, связанного с появлением новых источников огромного объема информации — как правило, неструктурированной.

Несмотря на популярность этой темы, по моим наблюдениям, существует недостаток качественной информации о ней. Если вы уже изучали ее, то наверняка сталкивались со множеством буклетов и статей, описывающих всю значимость больших данных, но не дающих никаких полезных деталей. Подозреваю, что они не показались вам убедительными и несущими практическую пользу. Возможно, вы находили статьи с подробным описанием различных технических терминов типа Nadoop, MapReduce и т. п. Но если вы не ИТ-специалист, то далеко не все из этого вам было интересно и понятно.

Книга, которую вы держите в руках, уникальна. На мой взгляд, это первая книга про большие данные, которая написана именно для бизнес-пользователей: руководителей разного уровня, аналитиков, маркетологов, экономистов. В ней прекрасно раскрыта тема больших данных с самых разных сторон: экономической, технологической и организационной. Вы найдете много полезной информации о необходимости изменений в вашей организации. Помимо новых технологий

вам потребуются правильные сотрудники с нужными компетенциями как для разработки аналитических идей по использованию больших данных, так и для реализации этих инициатив в жизни. Ведь важно, чтобы созданная в вашей компании аналитическая экосистема и культура аналитических инноваций способствовала не только накоплению больших объемов сырой информации, но и была нацелена прежде всего на извлечение из нее реальной пользы.

Я очень рад, что эту книгу написал сотрудник Teradata — компании, которая уже более тридцати лет занимается задачами хранения, обработки и анализа данных. У нас собрана уникальная команда, которая сфокусирована именно на этих задачах, и мы готовы делиться с вами своими знаниями и опытом. Используя передовые технологии, мы позволяем своим клиентам извлекать из данных нужные знания, встраивать их в операционные процессы и в конечном итоге конвертировать все это в прибыль. За кейсами, описанными в этой книге, стоят идеи и опыт применения наших решений крупнейшими передовыми мировыми компаниями. Они используют большие данные, бизнес-аналитику и инновации как свое конкурентное преимущество, поэтому остаются лидерами в своих индустриях.

Уверен, после прочтения этой книги у вас не только сложится правильное понимание, что такое большие данные, но и появится ряд практических идей по улучшению вашего бизнеса или компании, в которой вы работаете. Во-первых, вы поймете, что вы уже имеете в готовом виде, а именно какими данными, компетенциями сотрудников и технологиями вы располагаете. Во-вторых, сможете оценить, чего вам не хватает и что потребует изменений. Возможно, стоит подумать про реорганизацию существующих подразделений, оптимизацию некоторых бизнес-процессов и внедрение новых решений для определенных задач.

С большими данными ваш бизнес действительно может стать более конкурентоспособным, инновационным и потому экономически более эффективным! Но откладывать нельзя, нужно действовать уже сейчас. Большие данные никуда не исчезнут, они неизбежны, и игнорировать их нельзя. Ведь те, кто первым укротит большие данные и начнет правильно их использовать в своем бизнесе, будут иметь большое преимущество и серьезный лидерский отрыв в гонке с конкурентами. Удачи вам на этом пути!

*Андрей Алексеенко,
глава Teradata в России*

*Эта книга посвящается Стейси, Джесси и Даниэль.
Они мирились с тем, что многие ночи и выходные
я посвящал этой книге*

Предисловие

Хотите вы этого или нет, но в ближайшее время на вас обрушится огромное количество данных. Возможно, уже обрушилось. Возможно, вы уже на протяжении некоторого времени пытаетесь справиться с этим, понять, как хранить данные для последующего доступа, как исправлять ошибки и недостатки или классифицировать их. Теперь вы готовы извлечь смысл из этого огромного набора данных путем их анализа, чтобы узнать что-то о своих клиентах, своем бизнесе или о некоторых аспектах своей организационной среды. А возможно, вы пока далеки от этого, но уже видите свет в конце туннеля управления данными.

В любом случае вы пришли по адресу. Билл Фрэнкс предполагает, что вскоре мир наводнят не только большие данные, но и книги о больших данных. Я предсказываю (без всякой аналитики), что эта книга будет отличаться от прочих. Во-первых, она одна из первых на эту тему. Но, самое главное, она сконцентрирована на ином.

Большинство книг о больших данных будут посвящены управлению большими данными: тому, как собирать их в базу данных или хранилище данных, или тому, как структурировать и классифицировать их. Если вы много читаете о Hadoop, MapReduce или других методах хранения данных, это значит, что вы наткнулись на книгу, посвященную управлению большими данными.

Это, конечно, важная работа. Независимо от их объема и качества данные мало чем полезны, если их не поместить в такую среду и формат, которые позволят получить к ним доступ и проанализировать их.

Сама по себе тема управления большими данными не обеспечивает движения вперед. Для того чтобы извлечь пользу из данных, необходимо проанализировать их и совершить какое-либо действие

на основании результатов анализа. Так же как традиционные инструменты управления базами данных не обеспечивали автоматический анализ данных о транзакциях, полученных из традиционных систем, системы Hadoop и MapReduce не производят автоматическую интерпретацию данных, полученных от сайтов, картирования генов, анализа изображений или других источников больших данных. Даже до наступления эпохи больших данных многие организации многие годы (а иногда и десятилетия) занимались исключительно управлением данными, не извлекая из них никакой пользы в плане улучшения качества анализа и принятия решений.

Думаю, эта книга акцентирует внимание именно на том, на чем нужно. Она в первую очередь посвящена эффективному анализу больших объемов данных, а не управлению ими. Она начинается с данных и переходит к таким темам, как фреймовое представление решения, построение аналитического центра и создание аналитической культуры. Разумеется, здесь упоминается об управлении большими данными, однако основное внимание уделено созданию, организации, подбору персонала и воплощению аналитических инициатив, которые позволяют извлечь из входных данных пользу.

На тот случай, если вы этого не заметили: в настоящее время тема аналитики крайне актуальна в бизнес-среде. Я занимался в основном вопросами конкуренции компаний в области аналитики, и мои книги и статьи по этой теме были самыми популярными из всех, что я когда-либо писал. Конференции на тему аналитики проводятся повсеместно. У таких крупных консалтинговых фирм, как Accenture, Deloitte и IBM, имеется большой практический опыт в этой области. Многие компании, государственные и даже некоммерческие организации сделали аналитику своим стратегическим приоритетом. Сегодня наблюдается повышенный интерес к проблеме больших данных, однако в центре внимания должны по-прежнему оставаться способы приведения этих данных в форму, позволяющую проанализировать их и использовать в процессе принятия решений.

Билл Фрэнкс находится в уникальном положении: он может описать пересечение области больших данных и аналитики. Его компания Teradata, в отличие от других поставщиков систем хранения данных, всегда была максимально сосредоточена именно на анализе данных и извлечении из них пользы для бизнеса. И хотя компания

хорошо известна как поставщик корпоративных инструментов для хранения данных, Teradata в течение многих лет также предоставляла набор аналитических приложений.

За последние несколько лет Teradata наладила тесное партнерство с SAS — ведущим поставщиком аналитического программного обеспечения — для разработки высокомасштабируемых инструментов проведения анализа больших баз данных. Эти инструменты, которые часто подразумевают встроенный анализ в среде хранилища данных, предназначены для таких мощных аналитических приложений, как системы обнаружения мошенничества в режиме реального времени и крупномасштабного скоринга* покупательского поведения потребителей. Билл Фрэнкс — скоринг-директор по аналитике этого партнерства и поэтому имеет доступ к идеям и опыту в области проведения крупномасштабного анализа и «обработки в базе данных». Вероятно, лучшего источника на эту тему просто не существует.

Так что же еще особенно интересного и важного содержится в этой книге?

- ▶ Глава 1 включает в себя обзор концепции больших данных и объясняет, что «размер не всегда имеет значение». На протяжении всей книги Фрэнкс отмечает, что большая часть данных вообще бесполезна и очень важно уметь отфильтровывать ненужные данные.
- ▶ Обзор источников больших данных в главе 3 — интересный, полезный и необыкновенно подробный каталог. Подход к веб-данным и веб-аналитике в главе 2 может заинтересовать людей и организации, которые стремятся понять поведение потребителей, совершающих покупки через интернет. Этот подход выходит далеко за рамки обычной веб-аналитики, ориентированной на отчетность.
- ▶ Глава 4, посвященная «эволюции масштабируемости аналитических систем», представит вам технологические платформы для

* Скоринг (англ. score — подсчет очков) — система оценки кредитоспособности, в основу которой положены численные статистические методы обработки анкет потенциальных заемщиков. Суть ее в том, что за каждую позицию анкеты («стаж работы» или «количество детей») потенциальный заемщик получает некое количество баллов. В зависимости от суммы набранных баллов принимается решение об одобрении или отказе в выдаче кредита. *Прим. ред.*

больших данных и аналитики с такой точки зрения, которую вы больше нигде не найдете. В ней автор также описывает такие современные технологии, как MapReduce, и разумно утверждает, что анализ больших данных потребует использования комбинации сред.

- Эта книга содержит ультрасовременные сведения о том, как создавать аналитические среды и управлять ими, — эти сведения вы также нигде больше не найдете. Если вы хотите познакомиться с новейшими размышлениями на тему «аналитических песочниц» и «аналитических наборов данных предприятия» (это была новая для меня тема, однако теперь я знаю, что они собой представляют и какое значение имеют), вы найдете их в главе 5, которая также содержит важные замечания по поводу необходимости в системах и процессах управления моделями и скорингом.
- В главе 6 рассматриваются доступные сегодня типы аналитического программного обеспечения, в том числе программной среды R с открытым исходным кодом. Обычно очень трудно найти здравое рассуждение о сильных и слабых сторонах различных аналитических сред, однако здесь оно представлено. И наконец, описание методов анализа будет понятно даже далеким от техники людям.
- Третья часть книги сосредоточена на том, как управлять человеческим и организационным аспектами аналитики. В этом автор также опирается на здравый смысл. Мне, например, особенно понравился акцент на фреймовом представлении проблем и решений в главе 7. Слишком многие аналитики принимаются за анализ, не задумываясь о более важных вопросах, связанных с постановкой проблемы.
- Недавно меня спросили, описывал ли кто-нибудь, кроме меня, аналитическую культуру. Я сказал, что не знаю, однако это было до того, как я прочитал четвертую часть книги Фрэнкса. Она связывает аналитическую и инновационную культуру так, как никто прежде этого не делал.

Хотя книга содержит технические сведения, она доступна для широкой аудитории, в том числе для людей с ограниченными техническими

познаниями. Совет Фрэнкса по поводу инструментов для визуализации данных касается всей книги: «Чем проще, тем лучше. Прибегайте к усложнению только в случае крайней необходимости».

Если ваша организация собирается заняться аналитикой — а так и должно быть! — вам придется столкнуться со многими аспектами, затронутыми в этой книге. Даже если вы не специалист в технических вопросах, необходимо ознакомиться с некоторыми темами, связанными с обеспечением аналитических возможностей компании. Если же вы как раз являетесь техническим специалистом, то многое узнаете о человеческом аспекте аналитики. Если вы читаете это предисловие в книжном магазине или просматриваете описание книги на сайте, смело покупайте ее. Если вы ее уже купили, немедленно приступайте к чтению!

*Томас Дэвенпорт,
заслуженный профессор информатики и управления,
Бэбсон-колледж.
Сооснователь и директор по исследованиям
Международного института аналитики*

Введение

Вы получили электронное письмо: вам предлагают приобрести персонализированную компьютерную систему. Кажется, магазин прочитал ваши мысли, поскольку всего несколько часов назад вы просматривали информацию о компьютерах на его сайте...

Вы отправились в магазин за компьютером, и по пути поступает предложение купить со скидкой кофе в кофейне, мимо которой вы проезжаете: можете получить 10%-ную скидку, если заедете в течение ближайших 20 минут...

Пока пьете кофе, приходит извинение от производителя товара, на качество которого вы пожаловались вчера на своей странице в Facebook, а также на сайте компании...

Наконец, возвращаетесь домой, а вас ждет предложение приобрести специальную броню для вашей любимой онлайн-видеоигры, которая поможет пройти некоторые места, на которых вы застряли...

Звучит неправдоподобно? Думаете, это картины далекого будущего? Нет, эти сценарии возможны уже сегодня! Большие данные. Передовая аналитика. Аналитика больших данных. Кажется, что сегодня уже не обойтись без этих понятий. Люди обсуждают, пишут и продвигают идеи больших данных и передовой аналитики. Теперь к их суждениям можно добавить и эту книгу.

Что реально, а что нет? Уж слишком много внимания к этой теме! Может быть, анализ больших данных — не более чем шумиха? Разговоров на эту тему и правда много, однако эпоха преобразований в сфере аналитических возможностей и эффективного использования больших объемов данных действительно наступила. За ажиотажем, поднятым в средствах массовой информации, стоит нечто очень реальное

и мощное. Шумиха вокруг больших данных объясняется тем, что и предприятия, и потребители взволнованы ожиданием тех преимуществ, которые со временем предоставит анализ больших данных.

Большие данные, в свою очередь, становятся источником новых данных, которые стимулируют аналитические инновации в бизнесе, правительстве и академических кругах. Эти нововведения в состоянии радикально изменить взгляд организаций на свой бизнес. Большие данные обеспечат информацию, которая поможет принимать более взвешенные решения, и в некоторых случаях они будут разительно отличаться от тех, что принимаются сегодня. Анализ больших данных даст такое понимание, о котором сегодня можно только мечтать. Вы увидите, что укрощение волны больших данных и укрощение новых источников данных осуществляется аналогичными способами. Тем не менее дополнительные возможности, которые предоставляют большие данные, требуют использования новейших инструментов, технологий, методов и процессов. Старые способы анализа просто не сработают. Пришло время, когда передовые аналитические методы должны перейти на следующий уровень. Именно этому посвящена книга.

«Укрощение больших данных» не просто название книги. Скорее, это попытка определить, какие предприятия выиграют, а какие проиграют в следующем десятилетии. Подготовившись и взяв на себя инициативу, организации сумеют оседлать волну больших данных, чтобы достичь успеха, вместо того чтобы быть ею раздавленными. Что нужно знать и как подготовиться, чтобы подчинить себе большие данные и извлечь из них ценные новые сведения? Сядьте поудобнее и приготовьтесь это выяснить!

Целевая аудитория

В последние годы появилось бесчисленное количество книг, посвященных передовым методам анализа, а также ряд книг о больших данных. Эта книга подходит к вопросу с иной точки зрения. Основное внимание уделено объяснению, что такое большие данные и как с помощью аналитики их можно использовать, а также рассказать о подходах к созданию и развитию передовой аналитической экосистемы мирового класса в современной среде больших данных. Эта книга адресована широкому кругу читателей. Профессиональный ли вы аналитик,

предприниматель, использующий результаты работы аналитиков, или вам просто интересна тема больших данных — в этой книге вы найдете для себя что-нибудь полезное.

В книге нет подробных технических описаний; технические детали используются лишь в той мере, в какой необходимо обеспечить высокий уровень понимания обсуждаемой темы. Цель — помочь читателям понять и начать применять эти концепции, а также определить области для дальнейшего исследования. Эта книга скорее руководство, чем учебник, и она доступна для читателей, далеких от технических вопросов. В то же время те, кто уже глубоко понимает тему, между строк смогут увидеть технический подтекст.

Обзор содержания

Книга состоит из четырех частей, каждая из которых охватывает один аспект укрощения больших данных. В первой части объясняется, что такое большие данные, каково их значение и способы применения. Вторая часть касается инструментов, технологий и методов, необходимых для анализа и успешного использования больших данных. Третья часть посвящена людям, командам и принципам анализа, которые позволяют обеспечить эффективность. Четвертая часть подводит итог и фокусируется на том, как внедрить передовые методы анализа с помощью центра аналитических инноваций и изменения культуры. Приведем более подробное описание тем каждой части и главы.

Часть I. Появление больших данных

В первой части идет речь о том, что такое большие данные, почему они важны, в чем состоят преимущества их анализа. Описаны десять источников больших данных и то, как эти источники могут быть использованы организациями для улучшения своего бизнеса. Если читатели не знают, что такое большие данные или насколько широко их применение, первая часть даст ответы на эти вопросы.

Глава 1. Что такое «большие данные» и каково их значение? Эта глава начинается с обзора темы больших данных. Затем приводится ряд соображений о том, как организации могут их использовать. Для того чтобы помочь своим организациям справиться с волной больших

данных, читателям следует разобраться в содержимом данной главы так же хорошо, как в остальных главах.

Глава 2. Веб-данные: первые большие данные. Вероятно, наиболее широко используемый и самый известный источник больших данных на сегодняшний день — это данные, собранные с помощью сайтов. Журналы, которые содержат историю посещения пользователями веб-страниц, — настоящая сокровищница информации, которая только и ждет, чтобы ее проанализировали. Организации в целом ряде отраслей уже интегрировали подробные данные о клиентах, полученные с помощью сайтов, в собственную аналитическую среду. В этой главе показано, как эти данные расширяют возможности и изменяют процесс принятия различных бизнес-решений.

Глава 3. Источники больших данных и их ценность. Здесь мы подробно рассмотрим еще девять источников больших данных, чтобы объяснить, что представляет собой каждый источник данных, а также перечислим некоторые способы их применения в бизнесе. Одни и те же базовые технологии могут привести к возникновению нескольких источников больших данных в различных отраслях, а различные отрасли могут воспользоваться преимуществами одних и тех же источников данных. Большие данные имеют очень широкую сферу применения.

Часть II. Укрощение больших данных: технологии, процессы и методы

Часть II посвящена технологиям, процессам и методам, необходимым для укрощения больших данных. За последние годы увеличились возможности масштабируемости этих трех факторов. Организации не могут далее полагаться на устаревшие подходы и желают оставаться конкурентоспособными в мире больших данных. Эта часть книги наиболее «техническая», но все же она доступна для понимания. Читатели познакомятся с рядом концепций, с которыми им предстоит столкнуться в мире анализа больших данных.

Глава 4. Эволюция масштабируемости аналитических систем. Темп роста объема данных всегда предъявлял высокие требования к наиболее масштабируемому из доступных методов анализа. Перед появлением больших данных они уже были близки к своим пределам. Теперь традиционные подходы просто не работают. В этой главе

рассматриваются слияние аналитической среды со средой данных, массивно-параллельные архитектуры, облачные и грид-вычисления, а также модель MapReduce. Каждая из этих парадигм обеспечивает большую масштабируемость и будет играть важную роль в процессе анализа больших объемов данных.

Глава 5. Эволюция аналитических процессов. Значительное увеличение уровня масштабируемости требует обновления аналитических процессов. Глава начинается с описания использования так называемых аналитических песочниц для обеспечения профессиональных аналитиков масштабируемой средой в целях создания передовых аналитических процессов. Далее объясняется, как наборы данных предприятия могут обеспечить большую последовательность и уменьшить риск при создании аналитических данных и одновременном увеличении производительности труда аналитика. В конце главы описывается, как встроенные процессы скоринга позволяют пользователям и приложениям использовать результаты применения передовых аналитических процессов.

Глава 6. Эволюция аналитических инструментов и методов. В этой главе рассматриваются пути развития передовых аналитических инструментов, а также объясняется, как подобные прорывы повлияют на работу профессиональных аналитиков с большими объемами данных. Затрагиваются такие темы, как эволюция визуальных интерфейсов, аналитические точечные решения, инструменты с открытым исходным кодом и инструменты визуализации данных. Рассказывается, как профессиональные аналитики изменили свои подходы к построению моделей для более эффективного использования имеющихся возможностей. Среди описываемых тем: групповое моделирование, экспресс-моделирование и анализ текста.

Часть III. Укрощение больших данных: люди и подходы

Третья часть посвящена людям, которые занимаются анализом, их командам и подходам, используемым для обеспечения высокого качества работы. Наиболее важный фактор при проведении любого анализа, в том числе анализа больших данных, — наличие подходящих людей, которые руководствуются правильными принципами анализа.

Ознакомившись с третьей частью, читатели будут лучше понимать, чем хороший анализ, хороший профессиональный аналитик и хорошая команда аналитиков отличаются от остальных.

Глава 7. Что такое хороший анализ? Подсчет статистики, составление отчета и применение алгоритма моделирования — лишь некоторые из необходимых шагов для обеспечения хорошего анализа. В начале данной главы поясняются отдельные определения, а затем речь идет об обеспечении качественного анализа. Большие данные — довольно сложная тема, поэтому особенно важно понять принципы, излагаемые в этой главе.

Глава 8. Что такое хороший профессионал в области аналитики? Навыки в области математики, статистики и программирования — необходимые, но недостаточные характеристики хорошего профессионального аналитика. Хороший аналитик должен иметь такие качества, как обязательность, творчество, деловая смекалка, навыки проведения презентации и интуиция. В этой главе описано, почему каждая из этих черт имеет большое значение для профессионального аналитика и почему ими не стоит пренебрегать.

Глава 9. Что такое хорошая аналитическая команда? Как организации следует создавать и поддерживать команды аналитиков, чтобы обеспечить оптимальный эффект? Каким образом команды вписываются в организацию? Как они должны работать? Кто должен отвечать за создание передовой аналитики? Здесь затронуты часто встречающиеся проблемы и изложены принципы, которые необходимо иметь в виду при создании аналитической команды.

Часть IV. Объединение пройденного: аналитическая культура

В четвертой части изложены хорошо известные базовые принципы, которым должна следовать организация, чтобы успешно внедрять инновации, используя передовые средства анализа и большие данные. Поскольку это фундамент многих дисциплин, внимание сосредоточено на том, какое отношение данные принципы имеют к передовой аналитике в современной корпоративной среде. Описываемые концепции, вероятно, знакомы читателям в отличие от способов их применения к области передовой аналитики и больших данных.

Глава 10. Создание условий для внедрения инноваций в сфере аналитики. Глава начинается с обзора некоторых принципов, лежащих в основе успешного внедрения инноваций. Далее объясняется, как они применяются в мире больших данных и передовой аналитики, с помощью концепции центра аналитических инноваций. Цель состоит в том, чтобы показать читателям, как можно обеспечить внедрение аналитических инноваций и укрощение больших данных в своих организациях.

Глава 11. Создание культуры инноваций и открытий. Глава посвящена созданию культуры инноваций и открытий. Она написана легко и непринужденно и дает пищу для размышлений о том, что требуется для создания культуры, способной к инновационному анализу. Изложенные в главе принципы хорошо известны. Тем не менее их стоит еще раз проанализировать, а затем подумать о том, как их применить к большим данным и передовой аналитике.

ЧАСТЬ I

Появление больших данных

ГЛАВА 1

Что такое «большие данные» и каково их значение?

Пожалуй, ничто так сильно не повлияет на сферу передовой аналитики в ближайшие годы, как постоянное появление новых и мощных источников данных. Если говорить об анализе потребительского рынка, время, когда можно было полагаться исключительно на демографию и историю покупок, осталось в прошлом. Практически в каждой отрасли существует по крайней мере один совершенно новый источник данных, который в ближайшее время появится в интернете, если его еще там нет. Одни источники данных широко используются в различных отраслях промышленности, другие — в очень небольшом количестве отраслей или ниш. Многие из этих источников данных попадают под определение, которое вызывает в последнее время много шума: «большие данные».

Большие данные появляются везде, и их умелое применение окажется конкурентным преимуществом. Игнорирование больших данных опасно для организации, поскольку так можно отстать от конкурентов. Чтобы оставаться конкурентоспособными, крайне важно, чтобы организации активно анализировали эти новые источники данных и воспользовались содержащимися в них ценными сведениями. Профессиональным аналитикам предстоит много работы! Нелегко будет объединить большие данные со всеми остальными данными, которые в течение многих лет применялись для анализа.

В начале главы объясняется, что такое «большие данные». Далее приведены соображения о том, чем они могут быть полезны организации.

Что такое «большие данные»?

Однозначного определения понятия «большие данные» не существует, однако можно сослаться на два описания сути этой концепции, с которой согласится большинство людей. Первое определение предложил Мерв Адриан из компании Gartner* в статье для журнала Teradata Magazine в первом квартале 2011 года: «Большие данные — это данные, сбор, управление и обработку которых невозможно осуществить с помощью наиболее часто используемых аппаратных сред и программных инструментов в течение допустимого для пользователя времени»¹. Другое хорошее определение появилось в докладе McKinsey Global Institute** в мае 2011 года: «Большие данные — это наборы данных, размеры которых выходят за пределы возможностей по сбору, хранению, управлению и анализу, присущих обычному программному обеспечению базы данных»².

Из этих определений следует, что то, что считается большими данными, будет изменяться по мере развития технологий. То, что когда-то было «большими данными», или то, что считается «большими данными» сегодня, будет отличаться от «больших данных» завтрашнего дня. Некоторыхстораживает этот аспект понятия больших данных. Приведенные определения подразумевают, что суть больших данных может отличаться в зависимости от отрасли или даже организации, если существует значительная разница в возможностях инструментов и технологий. Мы обсудим это более подробно в этой главе в разделе «Сегодняшние большие данные отличаются от завтрашних больших данных».

В докладе McKinsey отмечены несколько интересных фактов, которые дают представление об объеме существующих сегодня данных.

- За \$600 сегодня можно купить диск, способный вместить всю музыку мира.
- Каждый месяц через сеть Facebook пользователи обмениваются 30 миллиардами фрагментов информации.
- В среднем компании пятнадцати из семнадцати отраслей промышленности Соединенных Штатов имеют больше информации, чем Библиотека Конгресса США³.

* Gartner — исследовательская и консалтинговая компания, специализирующаяся на рынках информационных технологий. *Прим. ред.*

** McKinsey Global Institute — американская глобальная консалтинговая фирма. *Прим. ред.*

Слово «большие» характеризует не только объем

Хотя понятие «большие данные» подразумевает наличие большого количества данных, оно не относится только к объему данных. Большие данные характеризуются возросшей скоростью их передачи, сложностью и разнообразием по сравнению с источниками данных прошлого.

Понятие «большие данные» подразумевает не только их объем. Согласно Gartner Group, слово «большие» относится и к некоторым другим характеристикам источника больших данных⁴. Это не только возросший объем, но и возросшая скорость передачи и разнообразие источников. Такие факторы, разумеется, усложняют работу с большими данными, поскольку вам приходится иметь дело не просто с большим количеством данных, а с тем, что они поступают к вам очень быстро, в сложных формах и из разнообразных источников.

Легко понять, почему большие данные сравнивают с приливной волной и почему ее приручение — настоящий вызов! Методы, процессы и системы анализа, внедренные в организациях, будут использоваться до предела, а возможно, и сверх предела. Необходимо разработать дополнительные методы и процессы анализа на базе обновленных технологий и методов для того, чтобы эффективно анализировать большие данные и действовать на основании полученных результатов. Мы коснемся всех этих тем в данной книге, чтобы продемонстрировать целесообразность укрощения больших данных.

Что важнее: «большие» или «данные»?

А теперь устроим небольшую викторину! Остановитесь на минуту и попробуйте ответить на следующий вопрос, прежде чем читать дальше: что является самым важным в понятии «большие данные»: 1) слово «большие», 2) слово «данные», 3) оба слова или 4) ни одно из них? Задумайтесь об этом на минуту и, определившись с ответом, переходите к следующему абзацу. Мысленно проиграйте музыку, которую включают в игре, пока участники думают.

Теперь проверим, правы ли вы. Правильный ответ — вариант 4). В термине «большие данные» ни одну из составных частей нельзя считать важнейшей. Важнее всего то, как организации используют большие данные. Анализ больших данных, производимый вашей организацией, в сочетании с действиями, предпринимаемыми для улучшения вашего бизнеса, — вот что имеет значение.

Наличие большого источника данных само по себе не является дополнительной ценностью. Возможно, ваши данные *больше*, чем мои. Кого это волнует? На самом деле наличие любого набора данных, вне зависимости от размера, само по себе не добавляет какой-либо ценности. Собранные, но не используемые данные имеют не большее значение, чем старый хлам, хранящийся на чердаке или в подвале. Данные не имеют значения до тех пор, пока не будут помещены в контекст и использованы. Мощь больших данных, как, впрочем, любого источника данных, заключается в том, что с ними делают. Как они анализируются? Какие действия предпринимаются на основании полученных результатов? Как эти данные используются для совершенствования бизнеса?

Вокруг больших данных поднята такая шумиха, что многие полагают: только благодаря большому объему, скорости передачи и разнообразию они важнее всех других. Это не так. Как мы увидим далее в этой главе (в разделе «Большая часть больших данных не имеет значения»), в больших данных доля бесполезного или малозначимого контента намного выше, чем в любом привычном источнике данных. Когда вы отберете действительно нужную вам информацию, источник больших данных может показаться вам не таким уж большим. Но это ничего не значит, поскольку после обработки данных их объем не имеет значения. Важно то, что вы будете делать с полученными результатами.

Дело не в объеме данных, а в способе их использования!

Значимость большим данным придает вовсе не то, что они большие, и даже не то, что они представляют собой данные. Важно то, как вы анализируете и применяете эти данные для развития своего бизнеса.

Что делает большие данные интересными для вас и вашей организации? Во все не то, что они «большие». Самое интересное связано с новыми мощными средствами их анализа. Об этом и поговорим.

Чем большие данные отличаются от традиционных данных?

Большие данные отличаются от традиционных данных рядом важных характеристик. Не каждый источник больших данных имеет все перечисленные особенности, однако большинству свойственно следующее.

Во-первых, большие данные часто автоматически генерируются машиной без участия человека. Традиционные источники данных всегда предполагают присутствие человека. Возьмем, к примеру, розничные или банковские транзакции, записи с содержанием телефонных звонков, доставку товаров или выставление счетов на оплату. Все эти действия подразумевают присутствие человека, который способствует созданию данных. Кто-то должен внести деньги, сделать покупку, позвонить по телефону, отправить посылку или сделать платеж. В каждом случае частью процесса создания новых данных остается человек, совершающий какие-либо действия. С большими данными дело обстоит иначе. Многие источники больших данных генерируются вообще без взаимодействия с человеком, например встроенный в двигатель датчик генерирует данные, даже если никто его об этом не просит.

Во-вторых, большие данные обычно соотносятся с совершенно новыми источниками данных. Это не просто расширение возможностей сбора существующих данных. Например, через интернет потребители могут взаимодействовать с банком или магазином, однако выполняемые ими операции принципиально не отличаются от традиционных. Они просто выполняют те же операции через другой канал. Организация может собрать данные о транзакциях, совершенных через интернет, однако они мало чем отличаются от транзакций, которые совершались раньше. Тем не менее сбор данных о поведении потребителей в процессе совершения транзакции предоставляет принципиально новую информацию, о которой мы подробно поговорим во второй главе.

Иногда большой объем данных может превратиться в нечто новое. Например, вы, вероятно, в течение многих лет каждый месяц вручную снимали показания счетчика электроэнергии. Можно ли считать, что интеллектуальный счетчик, фиксирующий показания каждые 15 минут, предоставляет те же самые данные? Или эта информация совершенно иного качества, открывающая возможности для проведения более глубокого анализа? Об этом речь пойдет в третьей главе.

В-третьих, многие источники больших данных не замыслились как дружелюбные к пользователю. Впрочем, некоторые из них вообще не замыслились! Возьмем, к примеру, текстовые потоки от сайта социальных медиа. Пользователей невозможно убедить соблюдать определенные правила грамматики, синтаксиса или лексические нормы. Когда люди публикуют запись, вы получаете то, что получаете. Работать с такими данными в лучшем случае трудно, а в худшем —

отвратительно. О текстовых данных говорится в главах 3 и 6. Большинство традиционных источников данных дружелюбны к пользователю. Например, системы для отслеживания транзакций предоставляют данные в понятной форме, что облегчает их загрузку и работу с ними. Частично это было продиктовано исторически сложившейся необходимостью в эффективном использовании пространства. Для избыточных данных просто не было места.

Большие данные бывают неприглядными

Традиционные источники данных с самого начала разрабатывались с учетом определенных требований. Каждый бит данных имел высокую ценность, иначе он не был бы учтен. Поскольку стоимость хранения данных стремится к нулю, источники больших данных, как правило, содержат все, что может быть использовано. Это означает, что при проведении анализа необходимо разбираться в огромном количестве хлама.

И, наконец, потоки больших данных далеко не всегда представляют собой особую ценность. Большая часть данных может быть вообще бесполезной. В журнале логов содержится как очень полезная информация, так и не имеющая ценности. Необходимо отсортировать мусор и извлечь ценные и релевантные фрагменты информации. Традиционные источники данных с самого начала разрабатывались так, чтобы содержать на 100% релевантные данные. Это было связано с ограничениями масштабируемости: включение в поток данных чего-то неважного слишком дорого обходилось. Мало того что записи данных были предопределены заранее — каждый фрагмент данных имел высокую ценность. С тех пор изменилось одно важное обстоятельство: мы более не ограничены объемом носителя. Это привело к тому, что большие данные по умолчанию включают всю возможную информацию, а позже приходится разбираться в том, что же из собранного имеет значение. Зато есть гарантия, что ничего не будет упущено, но усложняет процесс анализа больших данных.

В чем сходство между большими данными и традиционными данными?

Любая животрепещущая тема вызывает различные, порой взаимоисключающие толкования. Существует мнение, что большие данные в корне изменяют способы анализа и использования его результатов.

Однако если вдуматься, это не так. Это как раз тот случай, когда шума выходит за рамки реальности.

Ни для кого не новость, что большой объем больших данных создает проблемы масштабируемости. Большинство новых источников данных поначалу считались большими и сложными. Большие данные — это просто очередная волна новых данных, которая раздвигает существующие пределы. Аналитики смогли приручить прошлые источники данных с учетом существовавших в то время ограничений, и большие данные тоже будут приручены. В конце концов, аналитики в течение длительного времени находились в авангарде изучения новых источников данных. Так и будет продолжаться.

Кто первым начал анализировать данные о телефонных звонках в телекоммуникационных компаниях? Аналитики. На своей первой работе я проводил анализ данных, записанных на магнитные ленты. В то время казалось, что данных было огромное количество. Кто первым начал анализировать данные с мест продаж в розничных магазинах? Аналитики. Сначала анализ данных о сотнях тысяч товаров в тысячах магазинов считался огромной проблемой. Сегодня это не так.

Профессионалы в области аналитики, которые первыми начали работать с этими источниками, имели дело с тем, что в то время считалось немыслимо большими объемами данных. Им необходимо было найти способ анализа и использования данных с учетом существующих в то время ограничений. Многие сомневались в том, что это возможно, а некоторые даже ставили под сомнение ценность таких данных. Это очень похоже на то, что происходит с большими данными сегодня, не так ли?

Большие данные не повлияют ни на задачи, которые решают профессионалы в области аналитики, ни на причины, по которым они это делают. Даже для тех, кто сейчас называет себя не аналитиками, а учеными в области науки о данных, цели и задачи остаются прежними. Конечно, решаемые проблемы будут эволюционировать вместе с большими данными — так было всегда. Однако в конце концов аналитики и ученые будут просто изучать новые и немыслимо большие наборы данных, чтобы обнаружить ценные тенденции и модели, как они всегда это делали. В этой книге под термином «профессиональный аналитик» мы подразумеваем как традиционных аналитиков, так и ученых. Более подробно мы поговорим об этих специалистах в главах 7, 8 и 9. Сейчас важно понять, что задачи, связанные с большими данными, не так новы, как может показаться.

Вам нечего бояться

Во многих отношениях большие данные не создают для вашей организации новых проблем. Укращение новых источников больших данных, которые раздвигают существующие пределы масштабируемости, — постоянная тема в мире аналитики. Большие данные представляют собой просто новое поколение таких данных. Профессиональные аналитики хорошо разбираются в решении подобных задач. Если ваша организация справляется с существующими массивами информации, она справится и с большими данными.

Большие данные потребуют изменения тактик, которые используют в своей работе профессиональные аналитики. Для обеспечения более эффективной работы с большими данными к традиционным аналитическим средствам добавятся новые инструменты, методы и технологии. Для отбора ценных сведений из потоков больших данных будут разработаны сложные алгоритмы фильтрации; будут усовершенствованы процессы моделирования и прогнозирования. Более подробно это обсуждается в главах 4, 5 и 6.

Перечисленные тактические изменения коренным образом не меняют цели или сам процесс анализа. Большие данные, безусловно, будут способствовать внедрению новых и инновационных средств анализа, и это заставит аналитиков проявлять творческий подход к работе в пределах существующих ограничений в масштабируемости. Большие данные с течением времени продолжают увеличиваться в объеме. Тем не менее их использование на самом деле не сильно отличается от того, чем аналитики всегда занимались. Они готовы ответить на вызов.

Риски, связанные с большими данными

С большими данными связаны определенные риски. Так, например, организация может оказаться настолько перегруженной большими данными, что не будет способна на какой-либо прогресс. Ключевой момент здесь, как мы увидим в главе 8, — наличие нужных людей, которые не допустят этого. Вам нужны правильные люди, способные справиться с проблемами, которые возникают с появлением больших данных. Если такие специалисты есть, организации могут избежать пробуксовки в своем развитии.

Другой риск заключается в том, что расходы по сбору больших данных растут быстрее, чем возможности организации по их использо-

ванию. Избежать этой проблемы можно, лишь обеспечив соответствующий темп развития. Нет необходимости братья за все сразу и с завтрашнего дня собирать 100% информации, поступающей из каждого нового источника данных. Необходимо собирать и изучать образцы новых данных. С их помощью можно провести экспериментальный анализ, чтобы определить, что действительно важно в каждом источнике и как каждый из них может быть использован. Основываясь на этом, организация будет готова к проведению полномасштабного эффективного анализа источника данных.

Вероятно, самый серьезный риск, связанный с источниками больших данных, — это конфиденциальность. Если бы все люди были хорошими и честными, то нам не пришлось бы беспокоиться о конфиденциальности. Однако это не так. Нехорошими и нечестными бывают не только люди, но и компании. Существуют даже нехорошие и нечестные правительства. Вот поэтому большие данные могут доставить неприятности. Проблему конфиденциальности, связанную с большими данными, необходимо решать, иначе их потенциал невозможно реализовать полностью. Без надлежащего ограничения большие данные могут поднять такую волну протеста, что некоторые их источники будут полностью закрыты.

Не так давно стало известно, как несоблюдение безопасности привело к тому, что номера кредитных карт и правительственные документы были украдены и опубликованы в интернете. Не будет преувеличением сказать, что, если данные где-то хранятся, кто-то рано или поздно попытается их украсть. Как только злоумышленники получают к ним доступ, они будут их использовать в своих целях. Из-за непродуманной или ненадлежащим образом определенной политики конфиденциальности крупные организации сталкивались с проблемами: данные были использованы таким образом, который пользователи не понимали или не одобряли, и это вызывало негативную реакцию. По мере развития сферы больших данных должны развиваться сферы самостоятельного и правового регулирования их использования.

Наличие саморегулирования критически важно. Оно говорит о том, что отрасли не все равно. Участники рынка должны обеспечить саморегулирование и разработать правила, которых может придерживаться каждый. Такие правила обычно более эффективны и менее жестки, чем те, которые вводятся государственными органами, когда отрасль не может контролировать себя самостоятельно.

С большими данными связаны большие проблемы конфиденциальности

Принимая во внимание природу многих источников больших данных, нетрудно понять, что конфиденциальность представляет собой серьезную проблему. При наличии подобных объемов данных всегда найдутся нечестные люди, которые попытаются использовать их без вашего согласия или таким образом, который вам вряд ли понравится. Правила обработки, хранения и применения больших данных должны развиваться наряду с аналитическими возможностями. С самого начала пересмотрите подход вашей организации к вопросам конфиденциальности. Ваша позиция должна быть совершенно ясной и прозрачной.

Люди уже обеспокоены тем, как отслеживается история просмотра веб-страниц. Существуют опасения по поводу отслеживания местоположения пользователей с помощью приложений для мобильных телефонов и GPS-систем. Раз несанкционированное использование больших данных возможно, рано или поздно кто-нибудь попытается это сделать. Значит, необходимо предпринять шаги, чтобы этого не допустить. Организации должны четко объяснить, как они будут обеспечивать безопасность данных и как будут их использовать, если они хотят получить разрешение пользователей на их сбор и анализ.

Почему большие данные необходимо укротить

Многие организации пока мало используют большие данные. На ваше счастье, если вы до сих пор игнорировали большие данные, в 2012 году ваша организация не слишком отстала от остальных (если только вы не относитесь к таким отраслям, как электронная коммерция, — анализ больших данных уже стал неотъемлемой частью этой сферы). Однако скоро все изменится, поскольку развитие этого направления быстро набирает скорость. До сих пор большинство организаций упускали возможность оказаться впереди всех, и для многих из них это вполне нормально. В настоящее время еще есть шанс опередить остальных. Через несколько лет любая организация, которая не занимается анализом больших данных, безнадежно отстанет. Осваивать большие данные необходимо уже сейчас.

Нечасто компании удается воспользоваться совершенно новыми источниками данных, чтобы извлечь из них пользу для своего бизнеса, пока конкуренты не сделали то же самое. Такую возможность предоставляют сегодня большие данные. У вас есть шанс опередить своих конкурентов. В ближайшие годы мы увидим множество примеров того,

как с помощью анализа больших данных компании полностью трансформируют себя; как конкуренты были застигнуты врасплох и остались далеко позади. Речь идет не только о таких модных новых индустриях, как электронная коммерция. Уже сейчас в публикациях, на конференциях и в других источниках приводятся убедительные примеры прорыва, в том числе компаний, работающих в скучных, старых и тяжеловесных отраслях. Мы расскажем об этом в главах 2 и 3.

Время пришло!

Ваша организация должна начать процесс освоения больших данных уже сейчас. Пока что, если вы до сих пор игнорировали большие данные, то лишь упустили возможность быть в авангарде. Сегодня вы еще можете оказаться впереди всех. А если будете оставаться в стороне, через несколько лет окажетесь далеко позади. Если ваша организация уже занимается сбором данных и использует анализ в процессе принятия решений, то переход к большим данным не будет проблемой. Это просто расширение той деятельности, которой вы занимаетесь сегодня.

Фактически решение об использовании больших данных не должно стать проблемой. Большинство организаций уже подходят к сбору и анализу данных как к одной из основных частей своей стратегии. Хранилища данных, отчетность и анализ используются повсеместно. Если организация понимает, что данные представляют собой ценность, работа с большими данными будет лишь расширением ее деятельности. Не позволяйте скептикам убедить вас в том, что исследование больших данных не стоит затраченных усилий, или что их ценность еще не доказана, или что это слишком рискованно. Те же самые доводы помешали бы прогрессу, достигнутому за последние несколько десятилетий в области анализа данных. Обратите внимание сомневающихся на то, что работа с большими данными — это лишь продолжение того, что организация уже делает. Большие данные не представляют собой чего-то принципиально нового, и их не следует бояться.

Структура больших данных

В этой книге часто говорится о том, что данные могут быть структурированными, неструктурированными, полуструктурированными или даже мультиструктурированными. Большие данные нередко описываются как неструктурированные, а традиционные данные — как структурированные. Однако границы между ними не столь ясны, как можно

понять из названия. Рассмотрим три типа структуры данных с точки зрения неспециалиста. Технические детали выходят за рамки данной книги.

Большая часть традиционных источников данных — полностью структурированные. Это означает, что традиционные источники предоставляют данные в четко predetermined формате. Он не меняется день ото дня или в зависимости от обновления. В случае торговли акциями в первом поле может указываться дата в формате ДД/ММ/ГГГГ. Далее может идти 12-значный номер счета. Затем может быть указан символ акции, состоящий из трех-пяти знаков. И т. д. Каждый фрагмент используемой информации известен заранее, представлен в определенном формате и подчинен определенному порядку. Это облегчает работу.

Источники неструктурированных данных — а к ним относятся текстовые данные, видео- и аудиоданные — вы не можете контролировать. Вы получаете то, что получаете. Изображение подразумевает такой формат, при котором отдельные пикселы располагаются в строках, однако их взаимное расположение, определяющее то, что видит зритель, существенно различается в каждом конкретном случае. Приведенные примеры источников больших данных относятся к совершенно неструктурированным. Однако значительная часть данных относится к категории полуструктурированных.

Полуструктурированные данные подразумевают логическую схему и формат, который может быть понятным, но недружественным к пользователю. Иногда полуструктурированные данные называются мультиструктурированными. В потоке таких данных кроме ценных фрагментов информации может присутствовать множество ненужных и бесполезных данных. Чтение полуструктурированных данных с целью их анализа вовсе не так же просто, как файла определенного формата. Чтобы прочитать полуструктурированные данные, необходимо использовать сложные правила, которые динамически определяют, что следует делать после чтения каждого фрагмента информации.

Логи, собираемые в журнальных файлах, — прекрасный пример полуструктурированных данных. Они выглядят довольно уродливо, однако каждый фрагмент информации служит определенной цели. Служит ли любой из фрагментов журнала именно вашей цели — это совсем другой вопрос. На рис. 1.1 изображен пример необработанных данных интернет-журнала.

Данные интернет-журнала

```
96.255.90.50 - - [01/Jun/2010:05:28:07 + 0000] "GET / origin-
log.enquisite.com/d.js?id=a1a3af-
ly61645&referrer=http://www.google.com/search?hl=en&q=budget+planner&aq=5&aqi=g
10&agl=&og=budget+&gs_rfai=&location=https://money.strands.com/content/simple-
and-free-monthly-budget-planner&ua=Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 6.0;
SLCC1; .NET CLR 2.0.50727; .NET CLR 3.0.30618; .NET CLR 3/5/30729;
InfoPath.2)&ps=pgys63w0xgn102in8ms37wka8guxe74e&sc=cr1kto0wmxgik1wlr9p9weh
6xyx8g8sa&r=0.07550191624904945HTTP/1.1" 200 380 "-" Mozilla/4.0 (compatible;
MSIE 7.0; Windows NT 6.0; SLCC1; .NET CLR 2.0.50727; .NET CLR 3.0/30618; .NET CLR
3.5.30729; InfoPath.2)" "ac=bd76aad174480000679a044cfda00e005b130000"
```

Рис. 1.1. Пример необработанных данных интернет-журнала

Какую структуру имеют ваши большие данные?

Многие источники больших данных на самом деле являются полуструктурированными или мультиструктурированными, а не совсем неструктурированными. Такие данные подразумевают логическую схему, которая позволяет извлечь информацию для анализа. С ними просто сложнее работать, чем с традиционными источниками структурированных данных. Использование полуструктурированных данных требует дополнительного времени и усилий для того, чтобы определить наилучший способ их обработки.

Хотя на первый взгляд может показаться иначе, данные интернет-журнала подчинены определенной логике. В них присутствуют поля, разделители и значения, как и в структурированном источнике. При этом они не согласованы друг с другом и не представляют собой набор. Текст журнала, сгенерированный только что щелчком кнопкой мыши на сайте, может быть длиннее или короче, чем текст, сгенерированный щелчком кнопкой мыши на другой странице минуту назад. И все-таки необходимо понять, что полуструктурированные данные не лишены логики. Вполне возможно найти взаимосвязь между различными их фрагментами — просто это потребует больше усилий, чем в случае со структурированными данными.

Профессиональных аналитиков больше тревожат неструктурированные данные, чем полуструктурированные. Возможно, им придется побороться с полуструктурированными данными, чтобы подчинить их своей воле, но они это сделают. Они смогут привести полуструктурированные данные в хорошо структурированную форму и включить в свои аналитические процессы. По-настоящему неструктурированные данные приручить гораздо сложнее, и это будет оставаться головной болью для организаций по мере того, как они будут учиться справляться с полуструктурированными данными.

Исследование больших данных

Начать работу с большими данными несложно. Просто соберите их и поручите команде аналитиков вашей организации разобраться в том, чем они могут быть вам полезны. Для начала не понадобится обеспечивать постоянный поток данных. Все, что вам нужно, — это позволить аналитической команде применить свои инструменты и подходы к некоторому набору данных, чтобы они могли начать процесс исследования. Это именно то, чем занимаются аналитики и ученые в области науки о данных.

Существует старое правило: 70–80% времени уходит на сбор и подготовку данных и только 20–30% — на их анализ. В начале работы с большими данными можно ожидать более низких значений. Вероятно, в самом начале аналитики будут тратить 95%, если не все 100%, времени только на то, чтобы разобраться в источнике данных, прежде чем они смогут решить, как его следует анализировать.

Важно понимать, что это нормально. Выяснение того, что собой представляет источник данных, — важная часть процесса анализа. Это, может быть, и скучновато, однако итеративная загрузка данных*, изучение того, как они выглядят, а также настройка процесса загрузки с целью более точного извлечения нужных данных критически важны. Без выполнения этих действий невозможно перейти к самому процессу анализа.

Приносите пользу по ходу дела

Чтобы решить, как использовать источник больших данных на благо своего бизнеса, придется потратить немало усилий. Аналитики и их работодатели должны подумать, как обеспечить небольшие быстрые достижения. Это продемонстрирует организации прогресс и обеспечит поддержку дальнейших действий. Такие достижения могут генерировать солидную отдачу от инвестиций.

Процесс выявления ценных фрагментов больших данных и определение наилучшего способа их извлечения имеют решающее значение. Будьте готовы к тому, что на это понадобится время, и не расстраивайтесь, если его потребуется больше, чем вы ожидали. По мере изучения

* Итеративная загрузка данных (от англ. iteration — повторение) — выполнение загрузки данных параллельно с непрерывным анализом полученных результатов и корректировкой предыдущих этапов работы. *Прим. ред.*

новых источников больших данных специалисты и их работодатели должны искать способы достижения небольших и быстрых побед. Если вы обнаружите хоть что-то ценное, это поддержит заинтересованность людей и продемонстрирует прогресс. Например, кросс-функциональная команда не может приступить к делу, а год спустя утверждает, что по-прежнему не может ничего сделать с большими данными. Необходимо придумать хоть что-то, и сделать это нужно быстро.

Вот отличный пример. Европейский розничный магазин. Компания решила начать использовать подробные данные интернет-журналов. При создании сложных долгосрочных процессов сбора данных они сначала наладили несколько простых процессов для определения того, какие товары просматривает каждый посетитель. Информация о просмотренных страницах была использована в качестве основы для последующей кампании, в рамках которой каждому посетителю, покинувшему сайт без совершения покупки, высылалось электронное письмо. Это простое действие принесло организации значительную прибыль.

Далее компания наладила долгосрочный процесс сбора и загрузки веб-данных. Важно то, что они даже не начинали работу со всем потоком данных. Представьте, какую прибыль они получают в будущем, когда приступят к более глубокому анализу этих данных! Сотрудники организации, с самого начала увидев реальные достижения, сохраняют высокую мотивацию, поскольку они уже оценили мощь даже самого простого использования данных. А главное, дальнейшие усилия уже оплачены!

Большая часть больших данных не имеет значения

Дело в том, что большая часть больших данных вообще не имеет значения. Неожиданно, не так ли? Однако так быть не должно. Мы уже упоминали, что поток больших данных подразумевает большой объем, скорость передачи, разнообразие и сложность. Большая часть содержимого потока данных не будет отвечать поставленным целям, а некоторая его часть вообще не будет иметь какого-либо значения. Укрощение больших данных похоже не на закачку воды в бассейн, а скорее на питье воды из шланга: вы отхлебываете только то, что вам нужно, а остальному позволяете течь мимо.

В потоке больших данных есть информация, которая имеет долгосрочное стратегическое значение; некоторые данные пригодны только для немедленного и тактического использования, а часть данных вообще бесполезна. Самое главное в процессе укрощения больших данных — определить, какие фрагменты относятся к той или иной категории.

Примером могут служить метки радиочастотной идентификации (RFID), речь о которых пойдет в главе 3. Они размещаются на палетах с товарами в процессе их перевозки; если это дорогие товары, метками помечают каждый из них. Со временем станет правилом помечать метками отдельные товары. Сегодня в большинстве случаев это связано с большими затратами, поэтому метки ставятся на каждой палете. Такие метки упрощают процесс отслеживания местоположения палет, позволяют определить, где они загружаются, разгружаются и хранятся.

Представьте себе склад с десятками тысяч палет. На каждом из них находится RFID-метка. Каждые 10 секунд считывающие устройства опрашивают склад: «Кто здесь?» Каждая палета отвечает: «Я здесь». Посмотрим, как в этом случае можно использовать большие данные.

Палета прибывает сегодня и сообщает: «Это палета 123456789. Я здесь». Каждые 10 секунд в течение следующих трех недель, пока находится на складе, палета будет снова и снова сообщать: «Я здесь. Я здесь. Я здесь». По завершении каждого опроса следует проанализировать все ответы на предмет изменения статуса палеты. Таким образом, можно подтвердить то, что изменения были ожидаемыми, и принять меры, если палета неожиданно изменила статус.

После того как палета покинула склад, она больше не отвечает на запрос считывающего устройства. После подтверждения того, что отбытие палеты было ожидаемым, все промежуточные записи с ответом «Я здесь» не имеют значения. По-настоящему важны только дата и время появления палеты на складе, а также дата и время ее отбытия. Если между этими датами прошло три недели, то имеет смысл сохранить только две временные метки, связанные с прибытием и отбытием палеты. Ответы, полученные с интервалом в 10 секунд, говорящие: «Я здесь. Я здесь. Я здесь», не имеют какой-либо долгосрочной ценности, однако собрать их было необходимо. Необходимо было проанализировать каждый ответ в момент его создания, однако долгосрочной ценности они не имеют, поэтому их спокойно можно удалить после отбытия палеты.

Будьте готовы отбросить данные

Одна из главных задач при укрощении больших данных — определить фрагменты, которые имеют ценность. Большие данные содержат информацию, пригодную для долгосрочного стратегического применения; данные, которые могут использоваться в краткосрочной перспективе, а также данные, которые вообще ничего не значат. Удаление множества данных может показаться странным, однако при работе с большими данными это в порядке вещей. Вам потребуется время, чтобы к этому привыкнуть.

Если необработанные большие данные можно сохранить в течение некоторого периода, это позволит вернуться к ним и извлечь дополнительные данные, пропущенные при первоначальной обработке. Хороший пример такого подхода — процесс отслеживания веб-активности. Большинство сайтов используют метод, основанный на тегах: необходимо заранее определить текст, изображения или ссылки, взаимодействие пользователей с которыми требуется отслеживать. Теги, которые не видны пользователю, сообщают о его действиях. Поскольку данные поступают только об элементах, содержащих тег, большая часть информации не учитывается. Проблема может возникнуть, если по каким-то причинам не выполняется запрос на тегирование нового рекламного изображения, в результате чего упускается возможность проанализировать взаимодействие с ним. Это изображение должно быть помечено тегом, прежде чем пользователь его увидит. Можно добавить тег и позже, однако в этом случае собираться будут только данные, полученные после добавления тега.

Существуют новые методики, позволяющие проанализировать необработанные данные интернет-журналов и определить события, которые не были предопределены заранее. Эти методы основаны на использовании содержимого журнала, поскольку они опираются на непосредственно содержащиеся в них необработанные данные. Преимущество этих методов в том, что если вы забыли собрать данные о взаимодействии пользователей с рекламным изображением, то можете позднее вернуться и извлечь необходимую информацию. В этом случае ничего изначально не отбрасывается, а нужные данные определяются в процессе анализа. Это важное преимущество, и именно поэтому хранение некоторого объема архивных больших данных, если оно оправдано с экономической точки зрения, имеет смысл. Объем архивных данных зависит от размера канала и от доступного

пространства для хранения данных. Хорошая идея — хранить такой объем архивных данных, который экономически оправдан с учетом доступного объема хранилища.

Эффективная фильтрация больших данных

Самая большая трудность при работе с большими данными может заключаться не в анализе, а в процессе извлечения, преобразования и загрузки данных (ETL), который необходимо наладить перед проведением анализа. ETL — это процесс сбора необработанных данных, их чтения и получения полезных выходных данных. Сначала данные извлекаются (E, *extracted*) из соответствующего источника. Затем они преобразуются (T, *transformed*) путем агрегации, комбинирования и применения функций, чтобы обеспечить возможность их дальнейшего использования. И, наконец, данные загружаются (L, *loaded*) в среду для анализа данных. Это и есть ETL-процесс.

Вернемся к нашему примеру. Когда вы пьете воду из шланга, вам все равно, какая часть потока воды попадет в рот. В случае с большими данными, напротив, очень важно, какие части потока данных будут собраны. Сначала вам потребуется изучить весь поток данных, и только после этого можно отфильтровать нужные вам фрагменты информации. Вот почему процесс укрощения больших данных может занять так много времени.

Как попить из шланга

Работу с большими данными можно сравнить с попыткой попить из шланга. Большая часть данных будет пропущена, как и большая часть воды. Цель в том, чтобы отхлебнуть нужное количество данных из потока, а не выпить его полностью. Если вы сосредоточитесь на важных фрагментах данных, то работать с большими данными будет проще.

Аналитические процессы могут потребовать наличия фильтров, чтобы при получении данных отбросить часть информации. По мере обработки данных будут применяться и другие фильтры. Например, при работе с данными интернет-журнала можно отфильтровать информацию о версии браузера или операционной системы. Такие данные редко бывают нужны. Позднее в процессе обработки можно отфильтровать данные о конкретных страницах или действиях пользователя, которые можно исследовать для решения бизнес-задач.

Сложность правил и объем отфильтрованных или сохраненных на каждом этапе данных зависят от источника данных и бизнес-задачи. Для достижения успеха решающее значение имеют правильные процессы загрузки и фильтры. Традиционные структурированные данные не требуют таких усилий, поскольку они заранее исследованы и стандартизированы. Большие данные часто приходится исследовать и стандартизировать в процессе анализа.

Объединение больших данных с традиционными данными

Вероятно, наибольший интерес представляет даже не то, что большие данные могут сделать для вашего бизнеса сами по себе, а то, что они могут сделать для бизнеса в сочетании с другими данными организации.

Так, например, мощный источник данных — история посещения веб-страниц. Информация о важности потребителя для организации и о покупках, совершенных им ранее через различные каналы, повышает ценность веб-данных, если их поместить в более широкий контекст. Мы поговорим об этом подробнее в главе 2.

Для предприятия коммунального обслуживания чрезвычайно значимы данные интеллектуальных сетей (Smart Grid). Знание закономерностей, связанных с оплатой счетов, типов жилищ потребителей и других факторов, делает данные, полученные от интеллектуальных счетчиков, еще более ценными. Об этом говорится в главе 3.

Текст электронной переписки с отделом обслуживания клиентов также ценный источник данных. Знание подробных спецификаций обсуждаемых товаров, информации о продажах и дефектах повышает важность этих текстовых данных. Об этом речь пойдет в главах 3 и 6.

Своей популярностью хранилища данных предприятия (EDW) по большей части обязаны не тому, что они дают возможность централизовать многочисленные витрины данных в целях уменьшения затрат на оборудование и программное обеспечение. Хранилища данных создают ценность, так как с их помощью можно объединять различные источники данных, благодаря чему они дополняют друг друга. Хранилища данных позволяют совместно анализировать данные о потребителях и сотрудниках, поскольку они хранятся в одном и том же месте. Они больше не являются полностью разделенными. Например, правда ли,

что одни сотрудники приносят компании больше дохода, чем другие? Ответить на такие вопросы гораздо легче, если данные хранятся в одном месте. Добавление больших данных увеличивает масштаб решаемых проблем, поскольку все больше новых типов данных могут быть объединены для обеспечения новых точек зрения и контекстов.

Комбинируйте!

Потенциал больших данных раскрывается в полной мере при взаимодействии с другими данными корпорации. Если включить итоги анализа больших данных в более широкий контекст, количество и качество полученных результатов стремительно возрастут. Вот почему большие данные должны быть частью общей стратегии работы с данными, а не отдельной стратегией, созданной специально для них.

Крайне важно, чтобы разработанная организацией стратегия работы с большими данными не отличалась от стратегии работы с традиционными данными. Это не приведет к нужному результату. Большие данные и традиционные данные — части общей стратегии. Большие данные должны быть еще одной гранью корпоративной стратегии работы с данными. С самого начала необходимо продумать и спланировать не только процессы сбора и анализа больших данных, но и то, как их использовать в сочетании с другими корпоративными данными, а также в качестве компонента целостного подхода к корпоративным данным.

Потребность в стандартах

Будут ли большие данные по-прежнему характеризоваться невероятными форматами, неограниченными потоками и отсутствием определенности? Вряд ли. Со временем будут разработаны стандарты. Многие источники полуструктурированных данных удастся структурировать, отдельные организации подстроят свои потоки больших данных, чтобы их было легче анализировать. Но, что еще более важно, со временем произойдет переход к отраслевым стандартам. Хотя текстовые данные вроде электронных писем и комментариев в социальных медиа невозможно контролировать, можно стандартизировать подходы к интерпретации таких данных и использовать их для анализа. Это происходит уже сейчас.

Например, какие слова считать «хорошими», а какие — «плохими»? В каких контекстах не применяются правила по умолчанию? Какие

из электронных писем требуют исчерпывающего разбора и анализа, а какие — лишь минимальной обработки? Стандарты производства больших данных будут развиваться, как и стандарты их обработки и анализа. Подвергнутся стандартизации и входные, и выходные данные. В результате упростится жизнь тех, кому поручено их укрощать. На это потребуется время, и многие из разработанных стандартов будут представлять собой, скорее всего, набор общепринятых передовых практик, применяемых специалистами, а не формальные правила или политики, разработанные официальными организациями, занимающимися стандартизацией. Тем не менее стандартизация будет развиваться.

Стремитесь к максимально возможной стандартизации

С помощью стандартов вы можете значительно облегчить свою жизнь, хотя вам не удастся стандартизировать все аспекты больших данных. Текстовые данные, например электронное письмо, невозможно контролировать на входе, но можно стандартизировать подходы к интерпретации таких данных и использованию их при проведении анализа. Сосредоточьтесь не только на стандартизации входного потока, но и на стандартизации способов использования больших данных.

Организации, которые быстро включатся в работу с большими данными, смогут повлиять на процесс разработки стандартов и, следовательно, обеспечить удовлетворение собственных потребностей. Некоторые отрасли даже работают на опережение. Еще до появления возможности сбора данных предприятия коммунального обслуживания начали работу по определению параметров данных интеллектуальных сетей. Если формальные определения и руководящие принципы разработаны заранее, данными интеллектуальных сетей гораздо легче управлять, чем если бы каждое предприятие только что начало работать с данными собственными способами, не обсудив их заранее с другими представителями индустрии.

Сегодняшние большие данные отличаются от завтрашних больших данных

Как мы уже упоминали, принятые определения понятия «большие данные» неоднозначны, а единого и точного не существует. Это понятие определено в относительных терминах, связанных с существующей

технологией и источниками. В результате то, что считается большими данными в одной компании или отрасли, может не считаться таковыми в другой. Понятие «большие данные» для крупной компании электронной коммерции будет отличаться от того, что считает большими данными мелкий производитель.

Еще более важно, что со временем изменятся характеристики больших данных, поскольку инструменты и методы работы с ними будут развиваться наряду с увеличением размеров хранилищ необработанных данных и вычислительной мощности. Десять или двадцать лет назад файлы с демографическими данными о миллионах клиентов, содержащие сотни полей, считались огромными и трудноуправляемыми. Сегодня эти данные умещаются на флеш-накопителе и могут быть проанализированы на низкопроизводительном ноутбуке. Понятия о большом объеме, высокой скорости передачи, большом разнообразии и сложности будут меняться вместе с большими данными.

Понятие «большой» изменится

То, что сегодня считается большими данными, не будет считаться большими данными завтра, так же как данные, считавшиеся большими десять лет назад, не считаются таковыми сегодня. Большие данные будут продолжать развиваться. То, что невозможно или немыслимо сегодня с точки зрения объема данных, скорости передачи, разнообразия и сложности, в будущем станет в порядке вещей. Так было всегда и так будет продолжаться в эпоху больших данных.

Данные о транзакциях в отраслях розничной торговли, телекоммуникаций и банковского дела считались очень большими и трудноуправляемыми еще десять лет назад. Фактически в конце 1990-х годов во многих организациях такие данные не были широкодоступны для анализа и отчетности. Сегодня эти данные считаются необходимым и основным активом. Практически каждая компания вне зависимости от своего размера имеет к ним доступ.

То, что пугает нас сегодня, не будет казаться страшным через несколько лет. Через десять лет поток кликов может стать стандартным легкообрабатываемым источником данных. Активная обработка каждого электронного письма, переписки с отделом обслуживания клиентов, а также комментариев в социальных медиа может стать обычной практикой для большинства организаций. Ежесекундное отслеживание сотен метрик, может быть, уже не потребует большого труда.

Пока мы будем осваивать существующие сегодня потоки больших данных, появятся новые источники еще больших данных. Что они будут собой представлять? Этого не знает никто. Попробуем представить себе, каким образом довольно быстро существующие источники данных могут превратиться в источники еще больших данных.

- ▶ Представьте себе, что история просмотра веб-страниц включает данные о движениях мыши и глаз пользователя, что позволяет уловить каждую деталь процесса навигации, а не только отследить элементы, по которым пользователь щелкнул кнопкой мыши. Это совершенно новый масштаб больших данных.
- ▶ Представьте, что телеметрические данные видеоигры больше не ограничиваются нажатием кнопки или совершенным действием. Что они также включают движение глаз и тела игрока, а также расположение и статус каждого объекта в сцене, а не только тех объектов, с которыми происходит взаимодействие. Это уже происходит.
- ▶ Представьте себе, что RFID-метка находится на каждом отдельном товаре в каждом магазине, на оптовой базе и заводе. Представьте себе, что эти чипы собирают десятки метрик в секунду, например данные о температуре, влажности воздуха, скорости, ускорении, давлении и т. д. Такой объем данных сегодня сложно себе представить.
- ▶ Представьте себе, что существует возможность записи и перевода в текст каждого разговора с отделом обслуживания клиентов или с отделом продаж. Добавьте к этому все электронные письма, переписку в чатах и комментарии в социальных медиа и на сайтах отзывов. Теперь попробуйте разобраться и проанализировать весь этот текст. Ваша голова еще не взорвалась?

Дело в том, что большие данные никуда не исчезнут. То, что страшит нас сегодня, не будет столь пугающим через несколько лет, однако к тому времени появятся новые устрашающие источники данных. Организациям необходимо будет продолжать корректировать свои методы и цели, чтобы обеспечить возможность использовать данные по мере их развития. Тем не менее, прежде чем корректировать и обновлять методы работы с большими данными, вашей организации необходимо с чего-то начать.

Обзор главы

Самые важные уроки этой главы.

- Большие данные часто определяются как данные, сбор, управление и обработку которых невозможно произвести с помощью наиболее часто используемых аппаратных сред и программных инструментов в течение допустимого для пользователя времени.
- Данные можно считать «большими» не только с точки зрения объема, но и с точки зрения разнообразия, скорости передачи и сложности.
- Мощь больших данных заключается не в том, что они «большие», или в том, что они «данные», а в их анализе и действиях, которые вы предпринимаете на основе его результатов.
- Большие данные часто автоматически генерируются машиной, обычно в недружественном пользователю формате. Обычно сначала собирают все, что возможно, а потом производится попытка определить, что имеет значение.
- Большие данные — это просто очередная волна новых данных, расширяющих существующие пределы. С точки зрения анализа они не отличаются от прошлых источников данных, которые тоже было сложно обрабатывать, когда они только появились.
- Большие данные изменяют некоторые тактики и аналитические инструменты, которые используют профессионалы, но они коренным образом не изменяют причин проведения анализа или того, как определяется ценность аналитики.
- Многие источники больших данных полуструктурированы. Хотя полуструктурированные потоки данных могут показаться не очень привлекательными, в них присутствует определенная логика. Большие данные могут быть неструктурированными, а также структурированными, как традиционные источники данных.
- Самые большие риски, касающиеся больших данных, связаны с конфиденциальностью. По мере развития сферы больших данных потребуются введение как самостоятельного, так и правового регулирования.

- ▶ Укрощение больших данных заключается не в том, чтобы контролировать все данные. Это больше напоминает попытку попить воды из шланга. Нужно отобрать только важные фрагменты.
- ▶ Самое интересное заключается в том, чем именно большие данные в сочетании с другими данными могут помочь бизнесу.
- ▶ Большие данные и традиционные данные — части общей стратегии работы с данными. Не разрабатывайте отдельную стратегию для работы с большими данными.
- ▶ Сфера больших данных будет продолжать развиваться. То, что мы считаем устрашающим сегодня, через десять лет не будет нас волновать, однако мы заинтересуемся новыми источниками данных.

ГЛАВА 2

Веб-данные: первые большие данные

Представьте, как было бы замечательно иметь возможность не только проследить за действиями потребителя, но и понять его намерения, разобраться в процессе принятия решения о том, будет он совершать покупку или нет. Раньше разобраться в этом было фактически невозможно. Сегодня подобные задачи можно решить, используя детальные веб-данные. Об этом и пойдет речь.

Лучший способ понять, что представляют собой большие данные, — разобрать на конкретных примерах, что это такое и как с их помощью компания может извлекать прибыль. Пожалуй, ни один источник больших данных не используется сегодня столь же широко, как веб-данные. Мы посвятим им всю эту главу, чтобы глубже разобраться в этой теме и подробно рассмотреть способы их применения⁵. В главе 3 мы расскажем о девяти остальных важных источниках больших данных и о том, как такие данные могут быть использованы.

Некоторые организации в целом ряде отраслей интегрировали подробные данные о поведении покупателей, полученные с сайта, в свои аналитические среды. Однако большинство организаций по-прежнему ограничивают веб-интеграцию использованием данных об онлайн-транзакциях. Поставщики традиционных средств веб-аналитики обеспечивают оперативную отчетность по показателям кликабельности, источникам трафика и показателям, основанным только на веб-данных. Тем не менее подробные данные о поведении в интернете до сих пор не использовались за пределами веб-отчетности.

Ведущие компании доказали, что подробные веб-данные могут увеличивать прибыльность компании. В этой главе мы расскажем о том, что делают эти организации, почему они это делают и почему каждой организации сегодня следует рассмотреть возможность использования подобных приемов. Примеры весьма убедительны даже для тех, кто еще не думал об интеграции подробных данных, генерируемых потоком кликов, с другими данными.

Основная тема этой главы — не просто укрощение веб-данных. Организациям стоит сосредоточиться не на агрегированных метриках из различных источников данных, а на интеграции веб-данных со всей остальной релевантной информацией о своих потребителях. Использование такой информации в масштабируемой аналитической среде позволяет понять намерения и предпочтения покупателей, а также разобраться в процессе принятия решения о покупке. Сведения, которые можно получить из этого нового источника данных, помогут организации сделать огромный шаг вперед.

Как организация собирает, анализирует и использует эту обширную информацию, чтобы получить ценные сведения? Во-первых, мы расскажем, какие данные необходимо собрать и зачем. Далее покажем на примерах, что именно могут выявить эти данные. Наконец, поговорим о том, как можно преобразовать аналитические процессы путем интеграции веб-данных. Веб-данные — это один из источников больших данных, который уже используется многими организациями. Добавьте в этот список свою компанию!

Обзор веб-данных

Организации на протяжении многих лет говорят о 360-градусном обзоре своих потребителей. Время от времени та или иная организация заявляет, что она обеспечивает себе полный 360-градусный обзор. На деле это невозможно, поскольку такой обзор подразумевает, что вы знаете о своих клиентах буквально все. Когда кто-либо говорит о 360-градусном обзоре, подразумевается, что организация имеет полное представление о своих клиентах, какое только возможно с учетом доступных на данный момент технологий и данных. Тем не менее финишная черта постоянно отодвигается. Как только вам начинает казаться, что вы достигли конца пути, финишная черта снова удаляется от вас.

Несколько десятилетий назад компания считалась передовой, если у нее были имена и адреса клиентов и она могла добавить демографи-

ческую информацию к этим именам с помощью новейших на тот момент сторонних сервисов для улучшения качества данных. Со временем передовые компании начали добавлять к данным о клиентах такие простейшие метрики, как давность, частота и денежная ценность (RFM). С помощью этих метрик отслеживаются время совершения клиентом последней покупки (давность); то, как часто он совершал покупки (частота) и сколько он потратил (денежная ценность). Эти показатели RFM могут быть подсчитаны за прошлый год и, возможно, за все время, пока человек остается клиентом компании. За последние 10–15 лет практически все предприятия начали собирать и анализировать подробную историю транзакций своих клиентов. Это значительно расширило аналитические возможности и позволило получить намного более глубокое понимание поведения клиентов.

Обновите свой 360-градусный обзор

Многие организации все еще анализируют данные о клиентах исключительно на основе совершенных транзакций. Интеграция новых источников данных, например веб-данных, уже возможна и обещает принести огромную пользу тем, кто поторопится сделать это раньше других. Использует ли ваша организация для анализа своих клиентов все доступные сегодня возможности?

Многие организации все еще находятся на стадии анализа истории транзакций. Хотя этот вид анализа по-прежнему важен, многие компании ошибочно полагают, что он остается единственным возможным 360-градусным обзором клиентов. Сегодня организации должны собирать большие данные о своих потребителях из таких точек соприкосновения с клиентами, как веб-браузеры, киоски, мобильные приложения, социальные сети и многие другие.

В свое время данные о транзакциях кардинальным образом изменили глубину анализа. Новые источники данных выведут аналитику на новый уровень. Сегодняшние возможности хранения и обработки данных позволяют добиться успеха, и многие передовые компании уже доказали это.

Что вы упускаете?

Вы когда-нибудь задумывались о том, что произойдет, если собрать данные, генерируемые сайтом? Возможно, 95% посещений не приводят к созданию корзины. Из 5% лишь около половины, то есть 2,5%,

начинают процесс оформления заказа. И из этих 2,5% всего две трети, или 1,7%, на самом деле совершают покупку. Во многих случаях эти значения являются вполне реалистичными.

Это означает, что, если вы отслеживаете только транзакции, вы упускаете более 98% данных о посещениях сайта. Но, что еще важнее, теряется более высокий процент доступных данных. К каждой совершенной покупке могут иметь отношение десятки или сотни конкретных действий, совершенных на сайте. Эта информация должна быть собрана и проанализирована наряду с итоговыми данными о совершенной продаже.

Важно отметить, что речь не идет о веб-аналитике прошлых лет. Традиционная веб-аналитика сосредоточивается на агрегированном поведении, суммированном в среде, включающей только веб-данные. Ваша цель — выйти за рамки отчетов, содержащих сводные статистические данные, и объединить данные о поведении клиента с другими данными о клиенте, полученными из других каналов. Это больше чем простые отчеты о показателе кликабельности и сводки о просмотрах страниц.

Так же как показатель RFM является лишь малой частью информации, которую могут предоставить данные о транзакциях, традиционная веб-аналитика — лишь часть сведений, которые предоставляют веб-данные. Веб-данные — это удивительная новая область, которая меняет условия игры и тем самым способна в корне изменить понимание организациями своих клиентов, а следовательно, существенным образом повлиять на их бизнес.

Представьте себе возможности

Представьте себе, что вы знаете обо всем, что делают ваши клиенты в процессе их взаимодействия с вашей организацией. А что если бы вам было известно не только о том, что они купили, но и о том, что они думали в процессе принятия решения о покупке и какими ключевыми критериями руководствовались? Такое знание обеспечило бы новый уровень понимания ваших клиентов, а также новый уровень взаимодействия с ними. Это позволило бы вам быстрее удовлетворять их потребности.

- Предположим, вы занимаетесь розничной торговлей. Вы ходите рядом с клиентами и записываете данные о каждой полке,

мимо которой они проходят, о каждом предмете, на который смотрят, о каждом товаре, который берут в руки, о каждом товаре, который кладут в корзину, а затем вытаскивают и возвращают на полку. Представьте себе, что вы знаете, читают ли они информацию о пищевой ценности, смотрят ли на инструкции по стирке, читают ли рекламные брошюры, которые находятся на полках, обращают ли внимание на другую информацию, представленную в магазине.

- ▶ Положим, вы управляете банком. Вы знаете обо всех характеристиках кредитных карт, рассматриваемых вашими клиентами. Представьте себе, что вы имеете возможность понять, что заставило их принять решение: условия программы лояльности, процентные ставки или стоимость годового обслуживания. Предположите, что вы знаете о том, что они говорят о каждом продукте после его приобретения.
- ▶ Положим, вы управляете авиакомпанией. Вы знаете обо всех рейсах, информацию о которых изучил каждый из ваших клиентов, прежде чем выбрать окончательный маршрут. Знаете, что интересовало клиентов больше: цена или удобство. Представьте себе, что вы знаете обо всех рассмотренных направлениях и о том, когда ваши клиенты впервые заинтересовались ими.
- ▶ А теперь вы владелец телекоммуникационной компании. Вы можете идентифицировать каждую модель телефона, тарифный план и вспомогательное оборудование, которые рассмотрели клиенты, прежде чем принять окончательное решение. Представьте себе, что вам известно о том, как они попали на ваш сайт: с помощью поискового запроса «продлить договор» или «расторгнуть договор».

Конечно, хорошо было бы иметь информацию, указанную в предыдущем списке. Вы можете получить ее прямо сейчас, взяв на себя обязательство по ее сбору и подготовке к анализу. В каждой из перечисленных отраслей есть организации, которые уже это делают.

Принципиально новый источник информации

Изучая подробные сведения о поведении клиентов, вы можете не только узнать о том, что они покупают, но и получить дополнительные сведения, например представление о том, как они принимают решения.

Вы видите не только результат, но и весь процесс покупки. Это не просто очередной источник больших данных. Многие организации горячо приветствовали возможность интеграции данных о веб-транзакциях с данными о традиционных транзакциях. Однако веб-транзакция, по сути, является очередной записью с новым типом или местоположением. Детальные же сведения о поведении в интернете могут предоставить данные, аналогов которым не существует. Это принципиально новый источник информации.

Редкая возможность

Нечасто организация имеет возможность собрать принципиально новые данные. Детальные веб-данные такую возможность предоставляют. На сегодняшний день не существует источника данных, который обеспечивает такие же сведения, как веб-данные, кроме дорогостоящих опросов или исследований, охватывающих лишь небольшую часть клиентов.

Один из самых интересных аспектов веб-данных заключается в том, что они обеспечивают информацию о предпочтениях, намерениях и мотивах клиентов, которую можно получить только из непосредственной беседы или благодаря опросам. Почему клиенты предпочитают одно предложение другому? Возможно, организации думают, что они это знают. Однако они, скорее всего, обнаружат, что многие потребители принимают решения совершенно неожиданным для них образом.

Как только вам будут известны намерения, предпочтения и мотивы потребителей, у вас появятся новые способы общения с ними, обеспечения дальнейшего взаимодействия и повышения их лояльности. Однако самое интересное произойдет тогда, когда вы объедините веб-данные со всем тем, что получено благодаря прежнему 360-градусному обзору. Теперь этот обзор может быть дополнен новыми детальными сведениями о поведении потребителей в интернете.

Какие данные необходимо собирать?

По возможности необходимо собирать данные обо всех действиях потребителя в процессе взаимодействия с организацией. Это означает, что нужно сохранять подробную историю событий, относящихся к любой

точке взаимодействия с клиентом. В настоящее время к точкам взаимодействия относятся сайты, киоски, мобильные приложения и социальные медиа. Существует возможность сбора данных о широком спектре событий (табл. 2.1).

Табл. 2.1

Действия, данные о которых можно собрать

Покупки	Запрос помощи
Просмотр информации о товаре	Отправка ссылки
Пополнение корзины	Публикация комментария
Просмотр видео	Регистрация на вебинар
Загрузка файла	Выполнение поиска
Чтение/написание отзыва	И многое другое!

Эта глава посвящена веб-данным, однако эти же принципы применимы и к другим источникам, перечисленным в первом абзаце. Приводимые далее примеры ориентированы на сайт, однако имейте в виду, что те же концепции относятся ко всем точкам взаимодействия, данные из которых могут быть собраны.

Речь идет не только о веб-данных

Существует множество точек взаимодействия, к которым могут быть применены концепции, обсуждаемые в этой главе. К ним относятся, например, киоски и мобильные приложения. Не ограничивайтесь только веб-данными.

Как насчет конфиденциальности?

Конфиденциальность сегодня является большой проблемой, которая со временем может стать еще более серьезной. Необходимо внимательно подойти к решению о том, какие данные нужно собирать и как их использовать. Вы должны уважать не только формальные юридические ограничения, но и мнение ваших клиентов. Вряд ли организации пойдет на пользу, если она создаст программы, которые вызовут их неприятие. Конфиденциальность требует серьезного разговора, но это выходит за рамки темы данной книги. Тем не менее рассмотрим один

из способов решения проблемы конфиденциальности, который в то же время позволяет извлечь выгоду из анализа веб-данных.

Даже если придерживаться консервативного подхода, существует возможность извлечь огромную выгоду из веб-данных. Даже в том случае, когда организация не желает взаимодействовать с отдельными потребителями или привязывать информацию к персональным данным о клиентах, веб-данные имеют для нее ценность. Произвольный идентификационный номер не позволяет идентифицировать потребителя, но может быть сопоставлен с каждым отдельным клиентом на основе факта входа в систему, файла cookie или похожего фрагмента информации. Это создает, так сказать, «обезличенные» данные о потребителе. Хотя все данные, связанные с одним из этих идентификаторов, относятся к одному и тому же человеку, люди, которые занимаются анализом, не имеют возможности связать эти данные с реальным потребителем. При этом анализ все так же дает возможность выявить модели поведения, свойственные различным клиентам. Эти модели представляют собой большую ценность, и вы можете обнаружить их, не задумываясь о том, что сделал конкретный человек.

Преимущества обезличенного анализа потребителей

Самая большая ценность анализа потребителей заключается в том, что он позволяет выявить агрегированные модели. Если необходимо применить прямой маркетинг, потребуется всего лишь установить личность потребителя по имени или адресу. С помощью обезличенных данных о клиентах можно произвести высококачественный анализ. При таком подходе аналитики знают каждого отдельного клиента по произвольному номеру, по которому невозможно установить его личность. Не упускайте возможность воспользоваться преимуществами такого анализа!

Важны закономерности поведения потребителей, а не поведение конкретного клиента. В данном примере отдельные люди рассматриваются в качестве источника входных данных. Не нужно определять личность каждого отдельного человека! Современные технологии баз данных позволяют специалистам производить анализ и без этого, избегая тем самым многих проблем конфиденциальности. Многие организации в результате проведения такого анализа на самом деле устанавливают личности конкретных клиентов и обращаются к ним с предложением. Предполагается, что эти компании разработали и соблюдают политику конфиденциальности, предусматривающую для клиентов возможность отказаться от предоставления личной информации.

Что позволяют обнаружить веб-данные

Теперь, когда мы разобрались в том, что собой представляют веб-данные, поговорим о них более подробно. Существует ряд областей, в которых веб-данные могут помочь организациям лучше понять своих клиентов. Если не освоить этот источник больших данных, то получить такие ценные сведения будет очень трудно, а то и вообще невозможно. В этом разделе главы показано, что вы можете получить, используя веб-данные, а в итоговом разделе представлено детальное описание их применения.

Покупательское поведение

Хорошая отправная точка для изучения покупательского поведения — знать, как клиенты попадают на сайт. Какие поисковые системы они используют? Какие ключевые слова вводят при поиске? Воспользовались ли они созданной ранее закладкой? Профессиональные аналитики могут использовать эту информацию для поиска закономерностей, определяющих, как поисковые запросы, поисковые системы и ссылающиеся сайты влияют на показатели продаж. Обратите внимание на то, что аналитики могут отследить не только показатели продаж в пределах конкретной веб-сессии, но и проанализировать показатели, относящиеся к одному и тому же клиенту за определенный период. Эту информацию можно объединить с данными о продажах на сайте и кросс-канальным обзором покупательского поведения за определенный период. Именно в этом и заключается ценность.

Как только потребители оказались на сайте, изучите товары, информацию о которых они просматривают. Выясните, кто из посетителей просто посмотрел целевую страницу, посвященную товару, а затем покинул ее, а кто продолжил изучение информации. Кто из посетителей просмотрел дополнительные фотографии? Кто прочитал отзывы? Кто изучил подробные характеристики товара? Кто просмотрел информацию о доставке? Кто воспользовался другой содержащейся на сайте информацией? Например, определите, какие товары были выбраны для сравнительного обзора. Наконец, определите, какие продукты были добавлены в список пожеланий или в корзину, а также выясните, не были ли они позднее оттуда удалены.

Читайте мысли своих клиентов

Веб-данные уникальны тем, что позволяют получить представление о том, что потребители намерены купить, а также понять, как они принимают решение о покупке. Это позволяет прибегнуть к проактивным мерам и подтолкнуть клиентов к покупке, которую им еще только предстоит совершить. Сделайте им правильное предложение, и они подумают, что вы читаете их мысли.

Одна из очень перспективных возможностей использования веб-данных заключается в том, что вы можете определить наборы интересующих потребителей товаров еще до того, как они совершили покупку. Выйдите за рамки попыток продать клиенту что-нибудь с помощью предложения, предоставляемого ему после совершения покупки. Вместо этого сразу предложите им купить все товары, информацию о которых они просматривают.

Положим, потребитель изучает информацию о компьютерах, резервных дисках, принтерах и мониторах. Вполне вероятно, что он рассматривает возможность полного обновления компьютерной системы. Предложите ему приобрести набор интересующих его товаров. Не ждите, пока клиент приобретет компьютер, чтобы после этого озвучить обычный набор дополнительных устройств. Предложение приобрести персонализированный набор устройств до совершения клиентом покупки сработает лучше, чем предложение приобрести обычный набор оборудования после того, как он уже что-то купил.

Пути к покупке и предпочтения потребителей

Веб-данные позволяют определить, как потребители принимают решения о покупке, на основании их навигации по сайту, а также получить представление об их предпочтениях. Возьмем, к примеру, авиакомпанию. На основании данных о заказанных билетах авиакомпания может судить о предпочтениях клиентов. Например, за сколько времени до вылета был заказан билет? Место в каком классе было заказано? Длилась ли поездка в течение выходных дней или нет? Все эти сведения очень полезны, однако веб-данные предоставляют авиакомпании еще больше информации.

Авиакомпания может выявить клиентов, которые ценят удобство. Такие клиенты обычно начинают поиск с указания конкретных дат и выбора прямых рейсов. Они откажутся от наиболее удобного прямого рейса только в случае существования большой разницы в цене при

минимальной разнице в уровне комфорта. Допустим, клиент может сэкономить \$700, приземлившись в нью-йоркском аэропорту имени Джона Кеннеди, а не в аэропорту «Ла Гуардия». При этом время приземления отличается на 30 минут, а дополнительные расходы на такси составляют всего около \$20. В этом случае клиент, ориентированный на удобство, может решить, что экономия в \$700 стоит дополнительных хлопот. Однако если выгода составляет всего \$200, а время прибытия сдвигается на два часа, то такой клиент выберет наиболее удобный для себя вариант.

Авиакомпания также может выявить клиентов, которых больше всего интересует цена и которые готовы рассмотреть множество вариантов, чтобы выбрать наиболее выгодное предложение. Такие клиенты откажутся от самого дешевого варианта только в случае существования небольшой разницы в цене и огромной разницы в уровне комфорта. Например, клиент может вылететь в 10 часов утра за \$220 или в 6 часов утра за \$200. Дополнительные четыре часа сна стоят \$20, поэтому ориентированный на цену клиент согласится на более высокую цену, чтобы приобрести билет на более поздний рейс.

На основании поисковых закономерностей авиакомпания сможет определить, насколько данный клиент привязан к конкретному предложению или маршруту. Изучил ли он все доступные возможности, прежде чем сделать выбор? Или просмотрел только конкретные маршруты и заплатил столько, сколько требовалось? Например, студент во время весенних каникул может рассмотреть любые предложения и выбрать наиболее выгодное. С другой стороны, клиента, который регулярно летает к своей семье, будет интересовать только то место, где проживают его родственники.

Если вы знаете о том, что клиенты регулярно просматривают предложения, действующие для определенных направлений в выходные дни, то поймете, что для них наиболее важно. Некоторые клиенты принимают решение слетать к семье, когда им попадается хорошее предложение. Если они найдут такое предложение, то воспользуются им. Как только эта закономерность выявлена, авиакомпания может лучше предупреждать потребности клиентов.

Эти примеры показывают, какие бесценные сведения могут быть получены путем объединения информации об истории покупок с данными о закономерностях в процессе исследования и просмотра страниц. Разумеется, потребуется время на изменение аналитических

процессов таким образом, чтобы они учитывали эти закономерности. Однако после определения аспектов сайта, которые привлекают каждого отдельного потребителя, можно будет более эффективно обращаться к ним с сообщениями, которые отвечают их потребностям.

Исследовательское поведение

Зная о том, как потребители используют контент сайта, вы решите, как именно следует взаимодействовать с каждым отдельным клиентом, а также определите, как различные аспекты сайта влияют на показатели продаж. На основании изучения возможностей, которые потребители рассматривают на своем пути к покупке, делается вывод о том, что для них важно.

Рассмотрим интернет-магазин, занимающийся продажей фильмов. Если пользователи, прежде чем принять окончательное решение, рассматривают различные форматы видео, например стандартное, широкоформатное или видео высокой четкости, это означает, что они в принципе готовы приобрести любые варианты, даже если в конечном счете обычно останавливают свой выбор на одном и том же формате. После выявления закономерностей целесообразно изменить то, что они видят при посещении сайта, чтобы упростить и ускорить поиск предпочтительных для них возможностей. Потребителю, который просматривает множество вариантов, можно каждый раз показывать все доступные форматы. При этом не стоит заставлять клиента разбираться во всех вариантах, если известно, что он никогда ими не интересуется и всегда покупает видео в одном и том же формате.

А вот другой способ использовать веб-данные для понимания исследовательского поведения потребителей: определите, какие фрагменты информации, присутствующие на сайте, ценятся клиентской базой в целом и лучшими клиентами в частности. Как часто посетители просматривают превью, дополнительные фотографии или технические характеристики, прежде чем совершить покупку? Имейте в виду, что отслеживание всех сеансов в сочетании с другими данными о клиентах позволяет установить, что человек проводил исследование в один день, а покупку совершил в другой. Итоговое событие часто представляет собой целенаправленную веб-сессию, во время которой совершается покупка. История просмотра страниц дополняет картину. Возможно, малоиспользуемая функция сайта, вопрос об удалении

которой рассматривает организация, является любимой функцией большого количества посетителей. В этом случае данную функцию стоит оставить.

Сила исследования

Чтобы получить представление о том, как потребители изучают товары перед совершением покупки, не нужно проводить дорогостоящие маломасштабные опросы. Веб-данные могут предоставить подробные сведения о том, что является важным для каждого конкретного клиента, а также для всех клиентов в совокупности. Кроме всего, исключается риск того, что при проведении опроса человек скажет вам одно, а сделает другое. В данном случае вы увидите правду.

Или, например, обнаружилось, что многие потребители отказываются от какого-то товара после подробного изучения его характеристик, но не тогда, когда они вообще их не просматривали. Изучив страницу, вы понимаете, что описанию товара не хватает ясности или точности. Вы исправляете описание, и это может привести к увеличению объема продаж.

Чтение отзывов — отличный способ узнать о том, что важно для людей. Кто из потребителей ценит отзывы? Кто нет? Какие товары обычно теряют клиентов после того, как те прочитают отзывы? Отзывы могут способствовать или мешать продажам. Если вы выявили потребителей, которые обычно совершают покупку после прочтения отзывов, и заметили, что многие из них решают не покупать конкретный товар после знакомства с мнениями о нем, следует просмотреть эти отзывы. Возможно, существует ряд отрицательных откликов. В этом случае вы выясните, есть ли под ними основание, на что они обращают внимание, и определите, как нужно решать названные проблемы.

Определение важных для посетителей возможностей сайта и того, как каждый из них использует его для проведения исследования, поможет лучше адаптировать сайт к каждому отдельному посетителю. Для потребителей, которые всегда просматривают подробные характеристики товаров, стоит сделать так, чтобы эти характеристики появлялись сразу при просмотре информации о товаре. Для тех, кто всегда интересуется фотографиями, можно предоставить фотографии в полном размере вместо миниатюр. Смысл состоит в упрощении процесса исследования для посетителей, чтобы при необходимости они пришли к вам, а не к вашим конкурентам.

Обратная связь от потребителей

Обратная связь — наиболее ценная информация, которую потребители могут сообщить вам о товарах и услугах. Если клиенты готовы потратить время на написание отзыва, это демонстрирует их лояльное отношение к бренду. С помощью интеллектуального анализа текста, позволяющего определить тон, намерения и тему сообщений клиентов, мы получим более полную картину того, что для них по-настоящему важно.

Существуют ли потребители, которые регулярно публикуют отзывы о купленных ими товарах? Если эти отзывы в основном положительные и другие клиенты их читают, то разумно было бы предусмотреть для этих потребителей систему мотивации, чтобы они продолжали в том же духе. Аналогичным образом, путем анализа вопросов и комментариев в онлайн-чатах с отделом обслуживания клиентов, можно не просто в общих чертах понять, о чем спрашивают потребители, но и то, о чем спрашивает конкретный клиент. Если анализ показывает, что конкретного клиента интересуют определенные характеристики товара, обратите его внимание на другие товары со сходными качествами.

Активно ли пользуется данный потребитель сетью Facebook? Отслеживает ли ваши сообщения в сервисе Twitter? Изучение комментариев и вопросов, которые ваши клиенты публикуют на подобных площадках, позволит узнать о том, что им нравится или не нравится. Выявив очень активных клиентов, которые часто высказываются о вашей компании в социальных медиа, вы можете сделать их влиятельными послами бренда. Уделяйте таким клиентам повышенное внимание, которого они заслуживают, учитывая то влияние, которое они оказывают на ваш бренд. Обратите внимание на то, что влияние потребителей не всегда напрямую связано с их платежеспособностью. Потребитель, имеющий среднюю платежеспособность, может иметь в определенных кругах большой вес, поэтому имеет смысл оценивать его в первую очередь по этому критерию.

Веб-данные в действии

Знания организации о своих клиентах никогда не дают абсолютно объективной картины, поэтому всегда необходимо стараться корректировать ее на основании доступной информации. Если у вас есть лишь

частичная картина, то для получения полного представления, как правило, бывает достаточно экстраполировать имеющиеся данные. Однако существует вероятность, что отсутствующая информация может полностью изменить ваши представления. В тех случаях, когда отсутствующая информация не согласуется с предположениями, не исключен риск неоптимальных, а то и совершенно неправильных решений.

Таким образом, организациям следует собрать и проанализировать максимально возможное количество данных. Мы рассмотрели различные типы веб-данных и некоторые способы их применения. Теперь перейдем к некоторым конкретным примерам того, как организации могут использовать веб-данные для улучшения качества анализа, проведения новых видов анализа и совершенствования своего бизнеса.

Следующее наилучшее предложение

Очень часто маркетинговый анализ проводится с целью предсказать следующее наилучшее предложение для каждого клиента. Какой из всех доступных вариантов может быть предложен клиенту в качестве следующего, чтобы максимизировать шансы на успех? Наличие данных о поведении клиентов в интернете способно полностью изменить степень точности решения в этом вопросе.

Предположим, вы работаете в банке и имеете следующие сведения о клиенте по фамилии Смит:

- ▶ У него есть четыре счета: расчетный, сберегательный, кредитная карта и автомобильный кредит.
- ▶ Каждый месяц он осуществляет 5 вложений и 25 снятий средств.
- ▶ Он никогда не посещает отделение банка лично.
- ▶ Общий размер его сбережений составляет \$50 000.
- ▶ Общий размер его задолженности составляет \$15 000 с учетом кредитной карты и автомобильного кредита.

Какое предложение стоит отправить г-ну Смицу по электронной почте? Проанализировав имеющиеся данные, ему можно предложить более низкую процентную ставку по кредитной карте или приобретение сберегательного сертификата. Однако никто бы не подумал предложить ему ипотечный кредит, поскольку ничто не говорит о его

актуальности. Однако после анализа поведения г-на Смита в интернете выясняется несколько ключевых фактов:

- За прошлый месяц он пять раз просматривал ставки по ипотечным кредитам.
- Он изучал информацию о страховании домовладельцев.
- Он изучал информацию о страховании от наводнения.
- За прошлый месяц он дважды рассматривал варианты ипотечных кредитов (сравнивал кредиты с разными видами ставок и сроками).

Теперь гораздо проще решить, что следует обсудить с г-ном Смитом, не так ли?

Действуйте на опережение

Изучение истории просмотра страниц может изменить направление, которое в противном случае могло быть выбрано. Решение может быть основано на информации, которую пользователь недавно просматривал. Во многих случаях это могут быть сведения о продуктах или продуктовых линейках, которых он раньше не покупал. Как только веб-данные оповестят вас о потенциальной возможности, можно принимать меры по привлечению внимания клиента к новой продуктовой линейке.

Организациям бывает трудно определить степень лояльности своей клиентской базы. Веб-данные предоставляют подсказки о том, что интересует клиентов, и позволяют определить, по-прежнему ли они лояльны к вашему бренду. Рассмотрим пример с розничным магазином, продающим товары через такой же каталог, который имеет множество физических магазинов. Наряду с другими данными каталогизатор собирает следующие данные о каждом клиенте:

- Последние просмотренные товары.
- Последние товары, о которых потребитель оставил отзыв.
- История покупок.
- Маркетинговая кампания и реакция потребителя на нее.

Данные обобщаются и анализируются с целью определить, какие товары больше всего интересуют потребителя. Далее производится

корректировка содержания отправляемого ему каталога, а также его объема и имеющихся в нем предложений. Это кардинально изменяет проведение рекламных мероприятий по сравнению с традиционным подходом, что обеспечивают следующие результаты:

- ▶ Уменьшение количества рассылаемых материалов.
- ▶ Уменьшение количества страниц каталога, содержащих специальные предложения.
- ▶ Существенное увеличение общей прибыли.

Веб-данные помогают полностью пересмотреть предпринимаемые действия и повысить их эффективность.

Моделирование потерь

Компании, работающие в телекоммуникационной отрасли, потратили огромное количество времени и усилий на то, чтобы создать, улучшить и довести до совершенства модели «оттока» клиентов. Эти модели позволяют выделить тех клиентов, которые с большой вероятностью могут расторгнуть договор обслуживания, что позволяет компаниям предпринять меры, чтобы это предотвратить. Отток клиентов — серьезная проблема для всей отрасли, на карту поставлены огромные деньги, поэтому эти модели сильно влияют на показатель чистой прибыли.

Управление оттоком клиентов в полной мере помогает понять, как потребители используют ваши товары и услуги, а также как обеспечить прибыльность. Сегодня этот процесс можно усовершенствовать, поместив веб-данные в подходящий контекст. Г-жа Смит, клиент телекоммуникационной компании Provider 101, вводит в поисковой системе Google запрос: «Как расторгнуть договор с компанией Provider 101?». Затем она переходит по ссылке на страницу, где изложены правила расторжения договора, предусмотренные этой компанией. Насколько более ценны и актуальны по сравнению с другими данными эти сведения для создания моделей оттока и дальнейшего принятия мер по его предотвращению!

Трудно назвать более точный показатель намерения расторгнуть договор, чем сведения о том, что г-жа Смит искала соответствующую

информацию. Аналитики могли заметить снижение уровня ее потребления, а могли и не заметить. В любом случае на выявление такого изменения в модели использования ушли бы недели, а то и месяцы. Отследив действия г-жи Смит в интернете, компания Provider 101 может отреагировать быстрее и предотвратить потерю этого клиента.

Упущение возможности раннего выявления клиентов, изучающих способы аннулирования договоров, означает попытку вернуть их тогда, когда они уже приняли решение и, возможно, уже выбрали другого поставщика. В большинстве случаев бывает уже слишком поздно, и клиент потерян навсегда.

Моделирование отклика

Многие модели создаются для того, чтобы прогнозировать, какой выбор сделает клиент, когда ему предоставится такая возможность. Обычно это попытка предсказать, какие из клиентов совершат покупку, примут предложение или перейдут по ссылке, содержащейся в электронном письме. Для создания таких моделей часто используется так называемая логистическая регрессия. Эти модели, как правило, называются моделями откликов, или моделями склонности. Модель оттока клиентов из предыдущего примера относится к этому же классу. Основное отличие заключается в том, что цель модели оттока — прогнозирование наступления отрицательного события (отток), а не положительного (покупка или отклик).

При использовании модели отклика, или склонности, все клиенты анализируются и упорядочиваются в соответствии с вероятностью совершения определенного действия. Затем на основе этого рейтинга создаются соответствующие сегменты для обеспечения охвата клиентской базы. В теории каждому клиенту соответствует уникальное количество баллов. Однако на практике, поскольку большинство моделей предусматривает лишь небольшое количество переменных, многие клиенты получают одинаковое или почти одинаковое количество баллов. Особенно это касается тех, кто совершает покупки не очень часто или тратит не очень много денег. В этих случаях множество клиентов может оказаться в больших группах с очень похожим и очень небольшим количеством баллов.

С помощью веб-данных мы существенно усилим дифференциацию клиентов. Это особенно актуально для тех, кто совершает покупки

нечасто или на небольшие суммы: их рейтинг может быть значительно поднят благодаря веб-данным. Рассмотрим пример, в котором рейтинг четырех клиентов подсчитывается с помощью модели отклика, учитывающей несколько переменных. Все клиенты в этом примере имеют одинаковое количество баллов, так как представляют одну и ту же ценность для каждой переменной модели. Показатели гипотетические, поэтому вас не должно волновать, как они были получены. Далее приведены данные о четырех клиентах:

- ▶ Последняя покупка была совершена в течение последних 90 дней.
- ▶ За прошлый год было совершено шесть покупок.
- ▶ В общей сложности потрачено от \$200 до \$300.
- ▶ Домовладелец с уровнем общего семейного дохода от \$100 000 до \$150 000.
- ▶ Участник программы лояльности.
- ▶ В прошлом году купил товар из категории рекомендуемых.

В данном случае все клиенты имеют одно и то же количество баллов и одинаковую вероятность отклика. Предположим, что у всех по 0,62 балла. Любая маркетинговая программа, основанная на этой модели, будет обращаться с каждым из этих четырех клиентов одинаково: приведенная информация не содержит никаких данных, которые отличали бы их друг от друга!

Теперь добавим веб-данные и увидим, как сильно изменится эта картина. Посмотрите, какие ценные сведения обеспечивают веб-данные:

- ▶ Клиент 1 никогда не просматривал ваш сайт, поэтому его балл снижается до 0,54.
- ▶ Клиент 2 в течение последнего месяца просматривал категорию рекомендуемых товаров, поэтому его балл повышается до 0,67.
- ▶ Клиент 3 в течение последнего месяца просматривал информацию о конкретном рекомендуемом товаре, поэтому его балл повышается до 0,78.
- ▶ Клиент 4 просматривал информацию о конкретном рекомендуемом товаре трижды на прошлой неделе, один раз добавил его в корзину, покинул корзину, а позднее вернулся к изучению этого товара. Его балл повышается до 0,86.

Поведение в интернете позволяет нам идентифицировать пользователей, заинтересованных в товаре или даже имеющих намерение его приобрести, а благодаря этому мы лучше дифференцируем клиентов, которых невозможно было дифференцировать прежде. Теперь распространите пример с этими четырьмя клиентами на миллионы клиентов по нескольким каналам и посмотрите, к каким изменениям это приведет!

Директор по маркетингу американского розничного магазина, торгующего специализированными товарами, на вопрос о значении использования веб-данных ответил: «Это похоже на печатание денег!» Хорошая новость заключается в том, что очень легко построить модель с веб-данными и без веб-данных, чтобы проверить, к каким результатам приводит их использование. Вы практически ничем не рискуете, когда тестируете их влияние в среде вашей организации.

Сегментация клиентов

Веб-данные позволяют использовать целый ряд совершенно новых аналитических приемов. Один из них заключается в сегментации клиентов на основании типичных для них закономерностей просмотра страниц. Такая сегментация обеспечит иной взгляд на клиентов, отличающийся от традиционных демографических или основанных на данных о продажах схем сегментации, а значит, уникальное понимание и действия.

Рассмотрим сегмент под названием «Мечтатели», который был выделен исключительно на основании просмотренных пользователями веб-страниц. Мечтатели часто помещают товар в корзину, а потом покидают ее. Они неоднократно добавляют один и тот же товар и отказываются от него. Это прежде всего касается таких дорогостоящих товаров, как телевизор или компьютер. Сегмент таких потребителей определить совсем нетрудно. Что же можно сделать после его выявления?

Один из вариантов действий заключается в изучении товаров, от которых отказываются клиенты. Возможно, клиент просматривает информацию о высокотехнологичном телевизоре, который стоит довольно дорого. В прошлом вы замечали, что этот клиент часто сначала нацеливается на дорогостоящий товар, а в итоге покупает менее дорогой. Отправка электронного письма с предложением менее дорогих

товаров, обладающих многими из интересующих их функций, может подтолкнуть таких потребителей к покупке телевизора.

Веб-данные требуют новых методов анализа

Для сегментации клиентов используются различные источники данных, например данные о продажах, демографические данные и результаты опросов. Теперь можно сегментировать клиентов и на основании просматриваемых ими страниц. Это дает представление о стилях покупательского поведения и процессах принятия решения, а также позволяет расширить список критериев сегментации.

Другой вариант носит оперативный характер. Статистика случаев, когда оставляют корзины, может быть скорректирована с учетом сведений о сегменте «Мечтатели». Если посетитель покидает корзину, организации часто рассматривают это как отказ. Однако изучение истории посещения страниц может указать на то, что 10 подобных случаев относятся к одному и тому же посетителю, который часто отказывается от товаров. В результате количество случаев оставления корзины может быть уменьшено и все факты отказа от данного товара могут быть расценены как один отказ. Это обеспечит более точные данные об отказах. Когда вы скорректируете статистику с учетом всех таких клиентов, средний показатель отказов несколько улучшится по сравнению с первоначальным значением. Мало того что новые значения будут выглядеть лучше — они еще и будут более точно отражать реальное положение вещей.

Оценка эффективности рекламы

Более точная оценка результативности платного поиска и онлайн-рекламы — еще одно преимущество использования веб-данных о поведении пользователей. Традиционная веб-аналитика обеспечивает отчеты высокого уровня, содержащие такие показатели, как общее количество кликов, количество поисковых запросов, стоимость клика или показа, ключевые слова, генерирующие наибольшее количество кликов, а также статистику, относящуюся к местоположению объявления на странице. Однако эти метрики относятся к агрегированной статистике и определяются на основании отдельных сеансов просмотра. Контекст также обычно ограничивается веб-каналом.

Это означает, что все статистические данные основываются только на том, что произошло во время одного сеанса, который был результатом поиска или клика по объявлению. Как только посетитель покидает сайт и его веб-сессия завершается, он выходит за рамки анализа. Эта статистика не учитывает прошлые или будущие посещения. Включение данных о просмотре страниц и расширение обзора за счет других каналов позволяют более точно оценить результативность платного поиска и рекламы.

- Связаны ли посещения сайта, вызванные кликом по объявлению или поисковым запросом, с наиболее ценными или наименее ценными клиентами?
- Сколько продаж сгенерировала первоначальная сессия в течение дней или недель, последовавших за первым кликом посетителя?
- Есть ли среди ссылающихся сайтов такие, которые направляют к вам посетителей, которые возвращаются чаще и совершают больше покупок, чем посетители, перешедшие с других сайтов?
- Выявил ли кросс-канальный анализ, учитывающий действия, совершенные в других каналах, то, что множество продаж во втором канале было осуществлено после того, как интерес был сформирован с помощью объявления или поиска?

Рассмотрим пример с финансовым учреждением. Заявки на оформление кредитных карт присутствуют везде. Они приходят по почте, печатаются в журналах и доступны в интернете. Банк в нашем примере понимает, что факт просмотра рекламы — лишь часть общей картины. О результативности рекламы говорит то, что происходит после совершения первого клика.

Банк производит масштабный анализ, чтобы выйти за рамки исследования кликов во время первого сеанса; также проводится исследование данных о клиентах за определенный период, чтобы проанализировать факты заполнения заявления, запросы, поступающие в отдел обслуживания клиентов, факты выпуска карты, факты активации карты, а также начальные расходы по карте. Такой анализ рекламы, выходящий за пределы исследования кликов, дает возможность более точно оценить эффективность рекламы и обеспечивает более целесообразное распределение рекламных бюджетов.

Зачем ограничиваться только тем, что происходит в данный момент?

Результат одного сеанса, начавшегося с просмотра рекламы, клика по ссылке, содержащейся в электронном письме, или поиска, не дает полной картины. Многие клиенты вернутся позже, чтобы завершить то действие, которое они начали, возможно, даже в другом канале. Традиционная веб-аналитика не учитывает действия, которые производятся после первоначального сеанса или до него. Расширьте свои возможности, чтобы делать и то и другое.

Благодаря подробным веб-данным о потребителях можно понять, какие объявления, ключевые слова или ссылающиеся сайты генерируют «лучшие» клики, поскольку эти данные дают более полную картину, чем агрегированные результаты анализа первоначальных веб-сессий. Дополнительные сведения, получаемые путем масштабного кросс-канального анализа, охватывающего данные за определенный период, обеспечивают недоступный ранее обзор. Дальновидные организации сумеют воспользоваться новыми стратегиями, недоступными для компаний, применяющих традиционные аналитические приемы. Это обеспечит их явное конкурентное преимущество.

Обзор главы

Самые важные уроки этой главы.

- ▶ Интеграция подробных веб-данных о поведении потребителей может кардинально изменить понимание организациями своих клиентов.
- ▶ Так же как появление данных о транзакциях сделало анализ более мощным и глубоким, веб-данные позволят вывести аналитику на новый уровень.
- ▶ Существуют и другие точки соприкосновения с клиентами, которые можно проанализировать так же, как сайт, например киоски и приложения для мобильных телефонов. К ним применимы те же принципы.
- ▶ Все данные, которые могут быть собраны, должны быть собраны. К таким данным относятся просмотры страниц, производство поиска, загрузка файлов, а также любые другие действия на сайте.

- Соблюдение конфиденциальности — одна из основных проблем, связанных с веб-данными, поэтому следует с осторожностью формировать политику использования этих данных. Будучи утвержденной, эта политика должна строго соблюдаться.
- Ценные сведения может дать анализ обезличенных данных о клиентах, которым присваивается произвольный идентификационный номер. При этом ни аналитики, ни кто-либо другой не может определить личность каждого отдельного человека. В данном случае значение имеют только закономерности.
- Веб-данные позволяют подробно изучить покупательское поведение клиентов, их пути к покупке, исследовательское поведение и обратную связь. Это можно сравнить с чтением их мыслей.
- Веб-данные позволяют улучшить результаты в таких сферах, как следующее наилучшее предложение, моделирование потерь, моделирование отклика, сегментация клиентов, а также анализ эффективности платного поиска и онлайн-рекламы.
- Возможность опередить конкурентов скоро исчезнет. Начните работать с этим источником больших данных прямо сейчас!

ГЛАВА 3

Источники больших данных и их ценность

Не правда ли, было бы здорово получить на мобильный телефон сообщение о скидке на обед в ресторане, мимо которого вы проезжаете? Вам понравилось бы, если бы распорядитель казино дал вам \$20, которые забыл заплатить вам крупье? Представьте, что вы можете быстро найти игроков онлайн-игры, чей стиль игры соответствует вашему, потому что игра может сообщить вам о том, кто они. Хотели бы вы снизить тариф на страхование автомобиля? Все это возможно благодаря большим данным.

В главе 2 шла речь о веб-данных, которые представляют собой наиболее широко используемый и признанный источник больших данных. Однако существует множество других источников больших данных, и все они имеют собственные области применения. Далеко не все из них хорошо известны. В этой главе мы подробно рассмотрим еще девять источников больших данных и способы их использования с целью предоставить вводную информацию о том, что собой представляет каждый из них. Затем рассмотрим некоторые способы их применения и значение, которое каждый источник данных представляет для бизнеса.

В главах 2 и 3 вы не найдете списка лучших источников, поскольку никто не возьмет на себя смелость утверждать, что именно эти источники больших данных наиболее важны. Порядок, в котором они перечислены, также не определяет их ценности. Задача в том, чтобы читатель узнал о доступных типах больших данных, а также о том, какие аналитические методы эти данные позволяют применять. Каждому читателю следует выбрать для себя по крайней мере некоторые из них.

Одна из наметившихся тенденций показывает, как одни и те же базовые технологии способны привести к появлению нескольких источников больших данных в различных отраслях. Кроме того, различные отрасли могут использовать одни и те же источники больших данных. Применение больших данных не сводится к одному способу. Их возникновение будет иметь долгосрочные последствия.

Речь пойдет о следующих источниках больших данных:

- Автострахование: значение телематических данных.
- Разные отрасли: значение текстовых данных.
- Разные отрасли: значение данных о времени и местоположении.
- Розничная торговля и производство: значение данных радиочастотной идентификации (RFID).
- Коммунальные предприятия: значение данных, генерируемых интеллектуальными сетями.
- Игровая индустрия: значение данных отслеживания фишек.
- Промышленные двигатели и оборудование: значение данных, полученных от датчиков.
- Видеоигры: значение телеметрических данных.
- Телекоммуникации и другие отрасли: значение данных, полученных из социальных сетей.

Автострахование: значение телематических данных

В сфере автострахования телематике стали уделять серьезное внимание. Телематика предполагает помещение в машину датчика, или «черного ящика», для сбора информации о том, что происходит с автомобилем. В зависимости от конфигурации это устройство отслеживает любое количество показателей, например скорость, пройденное расстояние или факт резкого торможения. Телематические данные позволяют страховым компаниям лучше оценить уровни риска клиента и более точно выбрать страховой тариф. Если не принимать в расчет проблемы конфиденциальности и представить крайний случай применения таких данных, то телематическое устройство может отследить, куда и когда ездил автомобиль, с какой скоростью он двигался и какие из его функций использовались.

Телематика позволяет снизить страховые тарифы для большинства водителей и увеличить прибыль страховых компаний. Как эти данные могут одновременно понизить тарифы и увеличить прибыль? Дело в том, что страховщики назначают размер страховых взносов исходя из оценки рисков. Использование традиционных методов оценки риска на основе демографических данных и персональной истории дорожных происшествий обеспечивает только общую картину. Особенно трудно охарактеризовать водителей, не попадавших в ДТП.

Страховые компании должны исходить из худшего сценария, поэтому они распределяют клиентов по группам с разной степенью риска, а затем принимают в расчет самый высокий уровень риска из присущих конкретной группе. Чем больше подробностей известно страховым компаниям о клиентах и их рисках, тем уже будет диапазон исков и, соответственно, тем в меньшей степени наихудший сценарий повлияет на повышение тарифа. Вот так одновременно тарифы могут снижаться, а прибыль компании повышаться. Страховщики могут точнее оценивать риски и уменьшить изменчивость прогнозируемых выплат.

Существуют страховые компании, которые используют телематические данные для страхования клиентов по всему миру, и число таких компаний растет. Ранние версии программ собирают минимальное количество информации об автомобилях. К примеру, они не отслеживают все места, в которых побывала машина. Эти программы фиксируют пройденное автомобилем расстояние, в какое время суток он находится в дороге, имело ли место превышение скорости и часто ли происходило резкое торможение. Это базовые данные, не создающие угрозу конфиденциальности. Поскольку приватные персональные данные не собираются, эта технология может получить более широкое распространение. Те же самые принципы применимы и в сфере коммерческой грузоперевозки. Установить тарифы на страхование грузовиков гораздо проще, если страховщик обладает более конкретными данными об их использовании.

Сначала телематические данные будут использоваться в качестве инструмента, обеспечивающего более эффективное страхование автомобилей и грузовиков. Со временем телематические устройства могут появиться в большом количестве других транспортных средств, что приведет к появлению новых способов использования телематических данных. Уже сегодня в автомобилях появляются бортовые компьютеры, однако телематические устройства могут вывести такие

системы на совершенно новый уровень. Существуют очень интересные методы использования телематических данных. Рассмотрим некоторые из них.

Использование телематических данных

Распространение телематических данных делает возможным применение фантастических аналитических методов. Представьте, что в миллионах или в десятках миллионов автомобилей в вашей стране находятся телематические устройства. Сторонняя исследовательская фирма получает у клиентов разрешение на сбор очень подробных анонимных телематических данных. В отличие от ограниченных данных, собранных для целей страхования, информация в этом примере включает поминутные или посекундные сведения об изменении скорости, местоположения, направления и т. д.

Этот поток данных будет предоставлять информацию о тысячах автомобилей, стоящих в любой пробке в любой день. Исследователи будут знать, насколько быстро движется каждый автомобиль. Они поймут, где началось движение, где оно закончилось и сколько времени длилось. Это удивительная детальная картина транспортного потока. Представьте себе последствия в сфере изучения пробок и планирования дорожной системы!

Выходите за рамки задуманного

Богатство возможностей телематики являет собой пример использования больших данных таким способом, который не предусматривался изначально. Часто наиболее эффективные способы применения того или иного источника данных кардинально отличаются от задуманных. Постарайтесь рассмотреть альтернативные методы использования каждого источника больших данных, с которым вы сталкиваетесь.

Как только исследователи получают доступ к тысячам автомобилей в каждый час пик, каждый день, в каждом городе они смогут досконально разобраться в причинах возникновения пробок и их последствиях. Они ответят на такие вопросы:

- Какое влияние на дорожное движение оказывают шины?
- Что произойдет, если левый ряд будет заблокирован?

- ▶ Каковы последствия сбоя синхронизации работы светофоров?
- ▶ Какие из перекрестков регулируются неэффективно, даже если они регулируются так, как было задумано?
- ▶ Как быстро пробка на одной полосе распространяется на другие полосы?

Сегодня ответить на эти вопросы позволяет только целенаправленное и дорогостоящее тестирование. Можно поставить на конкретный участок дороги человека, который будет фиксировать нужную информацию. Или установить датчики для подсчета проезжающих мимо автомобилей. Или поставить видеокамеру. Однако высокие расходы, связанные с использованием этих методов, ограничивают область их применения.

Упомянутые телематические данные — мечта инженера транспортного планирования. Если телематические устройства получают распространение, то можно будет изучить любой населенный пункт, достаточно многочисленный для того, чтобы на дорогах образовывались пробки. Изменения дорог и систем управления, а также планов их создания дадут огромные преимущества всем. Телематика изначально задумывалась как механизм, облегчающий процесс определения страховых тарифов. Однако она может кардинально изменить управление системами автомагистралей и улучшить нашу жизнь, уменьшив уровень стресса, который мы испытываем, простаивая в пробках.

Разные отрасли: значение текстовых данных

Текст — один из самых мощных и широко используемых источников больших данных. Только представьте себе существующий совокупный объем текста! Есть электронные письма, текстовые сообщения, твиты, комментарии в социальных медиа, мгновенные сообщения, чаты и аудиозаписи, переведенные в текст. Текстовые данные — один из наименее структурированных источников данных. К счастью, на сегодняшний день уже многое сделано для того, чтобы освоить текстовые данные и использовать их для принятия более эффективных бизнес-решений.

Анализ текста обычно начинается с его разбора и осмысления различных слов, фраз и компонентов, из которых он состоит. Это может быть сделано путем простого подсчета частотности употребления или

с помощью более сложных методов. Существует дисциплина под названием «Обработка естественного языка», она часто используется в таких аналитических методах. Но это не является предметом обсуждения в данной книге. Инструменты для интеллектуального анализа текста существуют в качестве как компонентов основных аналитических систем, так и автономных приложений для анализа текста.

В основе одних инструментов для анализа текста лежит подход, при котором пользователи должны настроить программное обеспечение для идентификации интересующих их закономерностей. Другие инструменты используют машинное обучение и прочие алгоритмы, позволяющие отыскивать модели автоматически. Каждый подход имеет свои преимущества и недостатки, однако их обсуждение выходит за рамки этой книги. Мы сосредоточимся не на получении результатов, а на их использовании.

После разбора и классификации приступают к анализу. Результаты, полученные в процессе анализа текста, часто используются в качестве входных данных для других аналитических процессов. Например, после определения тона электронного письма клиента генерируется переменная, которая определяет тон заказчика как негативный или позитивный. Теперь этот тег — часть структурированных данных, которые можно использовать в качестве входных для аналитического процесса. Создание структурированных данных на основе неструктурированного текста часто называется извлечением информации.

В качестве другого примера предположим, что мы знаем, о каких товарах клиент оставил комментарии в процессе общения с нашей компанией. Мы создаем набор переменных, которые определяют товары, обсуждаемые клиентом. Эти переменные также представляют собой структурированные метрики, которые можно использовать в процессе анализа. Эти примеры показывают способы сбора фрагментов неструктурированных данных и создания из них релевантных и структурированных данных.

Создавайте структуру там, где ее нет

Анализ текста — отличный пример того, как абсолютно неструктурированные данные могут быть обработаны и превращены в структурированные, которые используются в традиционных аналитических процессах. Один из основных аспектов процесса укрощения больших данных заключается в применении творческого подхода к процессу подготовки неструктурированных и полуструктурированных данных к дальнейшему использованию.

Интерпретация текстовых данных на самом деле довольно сложна. Смысл наших слов меняется в зависимости от того, какое из них мы акцентируем, а также от контекста, в который мы их помещаем. При взгляде на простой текст вы наверняка не знаете, на каком слове сделан акцент, и вам часто неизвестен весь контекст. Это означает, что придется сделать некоторые предположения. Мы поговорим об этом более подробно в главе 6.

Анализ текста — это одновременно искусство и наука, и он всегда будет подразумевать некоторый уровень неопределенности. При проведении анализа текста будут возникать проблемы, вызванные ошибками классификации и неоднозначностью. Это нормально. Если найденная в тексте закономерность позволяет принять более эффективное решение, то ее следует использовать. Цель анализа текста — улучшить принимаемые решения, а не достичь совершенства. Текстовые данные позволяют повысить качество принимаемых решений и предоставляют более ценную информацию, даже несмотря на содержащийся в них шум и неоднозначность.

Использование текстовых данных

Один из самых популярных вариантов анализа текста на сегодняшний день — исследование настроения. Анализ настроения позволяет изучить общее мнение большого количества людей, чтобы понять, о чем говорит рынок, что он чувствует и думает об организации. При этом часто используются данные социальных сетей. Вот некоторые примеры:

- ▶ В чем суть шумихи вокруг компании или продукта?
- ▶ О каких корпоративных инициативах говорят люди?
- ▶ Положительно или отрицательно высказываются люди об организации и ее товарах и услугах?

Мы уже говорили о том, что одна из сложностей анализа текста заключается в том, что слова могут иметь позитивное или негативное значение в зависимости от контекста. Это необходимо принимать во внимание, однако общее настроение множества людей должно быть ясно. Зная о том, что говорят люди в социальных сетях или при общении с отделом обслуживания клиентов, можно более уверенно планировать дальнейшие действия.

Если организация уловит настроения отдельного клиента, она сможет судить о его намерениях и мнениях. Подобно веб-данным, которые помогают определить намерения, мнение потребителя о товаре является ценной информацией. Это особенно верно, если потребитель ранее не покупал этот продукт. Анализ настроений показывает, насколько легко или трудно будет убедить клиента приобрести данный продукт.

Текстовые данные применяются для распознавания закономерностей. Анализируя жалобы, заявки на ремонт и другие комментарии, сделанные клиентами, организация сможет быстрее выявлять и решать вопросы, пока они не превратились в серьезные проблемы. После вывода нового продукта на рынок и начала поступления жалоб анализ текста поможет определить, с какими трудностями сталкиваются клиенты. Иногда удастся даже выявить назревающую проблему и предотвратить волну звонков в отдел обслуживания клиентов. Это позволит реагировать намного быстрее. Организация не только исправит дефекты в продуктах, которые будут выпущены позже, но и поможет клиентам справиться со сложностями, которые они испытывают сегодня.

Важной сферой использования текстовых данных является обнаружение мошенничества. В США в области страхования здоровья или трудоспособности, например, анализ текста можно использовать для разбора комментариев клиентов и заявлений на получение страховой выплаты. Затем выявляются закономерности, связанные со случаями мошенничества, чтобы оценить степень риска того или иного заявления. Заявления, которым присущ более высокий риск, следует проверять более тщательно. С другой стороны, некоторые заявления можно проверять автоматически. При наличии в заявлении закономерностей, терминов и фраз, относящихся к оправданным требованиям, его рассматривают как низкорисковое и проводят через систему в ускоренном режиме, а основные ресурсы сосредотачивают на заявлениях с более высоким риском.

Преимущества анализа текста используются и в сфере юриспруденции. В судебных делах часто анализируются электронная переписка и другие истории сообщений с целью выявить информацию, которая может иметь отношение к делу. Например, какие из электронных писем могут содержать инсайдерскую информацию? Кто предоставлял заведомо ложную информацию при взаимодействии с другими людьми? Что особенного в природе угроз?

Применение такого способа анализа в судебном делопроизводстве называют обнаружением электронных данных (eDiscovery). Все перечисленные методы анализа могут помочь в раскрытии преступлений. Без анализа текста, «вручную», было бы практически невозможно проверить все необходимые документы. Даже если такая попытка была бы предпринята, высока вероятность упустить ключевую информацию в связи с монотонностью занятия.

Текстовые данные могут быть востребованы во всех отраслях. Это будет один из наиболее широко используемых источников больших данных. Умение собирать, разбирать и анализировать текст имеет для организаций решающее значение. Текст — это один из источников больших данных, которые необходимо укротить.

Разные отрасли: значение данных о времени и местоположении

С появлением систем глобального позиционирования (GPS), персональных GPS-устройств и сотовых телефонов информация о времени и местоположении превратилась в постоянно растущий источник данных. Множество сервисов и приложений, таких как Foursquare, Google Places и Facebook Places, регистрируют местонахождение человека в каждый момент. Приложения сотовых телефонов могут отслеживать ваши местоположение и передвижения по вашему требованию. Даже при отсутствии функции GPS сотовые телефоны достаточно точно определяют местоположение, используя сигналы базовых станций операторов мобильной связи.

Существуют новейшие возможности использования этой информации потребительскими приложениями, пользователи которых решают собирать эти данные. Например, некоторые приложения позволяют отслеживать точные маршруты, которые вы проходите, когда занимаетесь спортом, их расстояние и время, которое вам требуется на их преодоление. Дело в том, что при наличии сотового телефона вы можете собрать данные обо всех местах, в которых побывали, и при желании предоставить эту информацию другим людям. Чем больше людей начинают обнародовать данные о своем местоположении, тем больше появляется интересных возможностей их использования.

Многие организации начинают понимать ценность знания о том, где и когда находятся их потребители, и стараются получить у них

разрешение на сбор такой информации. Разумеется, это всегда должно делаться на добровольной основе; кроме того, необходимо разработать четкую политику конфиденциальности и строго ее придерживаться. Сегодня организации придумывают привлекательные предложения, чтобы убедить клиентов предоставить им данные о том, где и когда они находятся.

Актуальны не только данные о местонахождении потребителей. Владелец грузовиков хочет знать, где находится каждый из них в любой момент. Владелец пиццерии интересуется, где сейчас находится каждый из разносчиков пиццы. Владелец домашних животных нужно знать, где находятся их питомцы, когда их выпускают из дома. Фирма, занимающаяся организацией банкетов, хочет знать, насколько эффективно обслуживаются клиенты.

Как только организация начинает собирать данные о времени и месте нахождения отдельных людей и предметов, она оказывается в области больших данных. Это особенно верно, если такая информация часто обновляется. Одно дело знать, где находится каждый из грузовиков в начале и в конце каждого дня, и совсем другое — знать, где он находится в каждый момент. Данные о времени и месте, а также способы их использования будут находить все большее применение.

Использование данных о времени и месте нахождения

Данные о времени и месте нахождения — очень спорный тип больших данных. Здесь возникают серьезные вопросы, которые связаны не только с конфиденциальностью, но и с этическими и моральными соображениями. Можно ли вживлять детям чипы, чтобы обеспечить возможность их найти в случае, если они пропадут без вести? А как насчет пожилых людей, страдающих слабоумием, которые уходят из своего дома или специализированного учреждения? Разумеется, существует вероятность злоупотребления данными о времени и месте нахождения. Однако их ценность при использовании надлежащим образом также высока. Рассмотрим несколько примеров.

Скоро люди смогут зарегистрироваться в местном отделении полиции или пожарной охраны и предоставить информацию о своем обычном маршруте передвижения. В случае какой-либо крупной аварии, наводнения, пожара или затора на дороге люди получают оповещение от противопожарной службы или полиции о том, что в определенном

месте их пути возникла нештатная ситуация, поэтому им следует воспользоваться другим маршрутом. Это ускорит дорожное движение. Со временем, если вы позволите, агентства смогут получать информацию о вашем местоположении в реальном времени.

На основе таких данных совсем недавно стали появляться предложения, учитывающие время и местоположение клиента. В будущем такие предложения приобретут огромную популярность в области маркетинга. Дело уже не просто в том, чтобы решить, что следует предложить клиенту сегодня или на этой неделе, а в том, чтобы сделать это исходя из того, где и когда клиент находится. Сегодня это, как правило, возможно после того, как клиент зайдет в систему и сообщит о своем местоположении. Когда-нибудь организации смогут сами отслеживать местонахождение клиентов постоянно и обращаться к ним по мере необходимости.

Например, клиент сообщает, что он будет находиться в пути с работы домой в 17:30 и проедет мимо Exit 5 между 17:45 и 18:00. Он собирается поужинать и хочет знать, что вы можете ему предложить, если он заедет в ваш магазин или ресторан. Вы должны предоставить ему то, что соответствует его потребностям, в тот самый момент и в том самом месте. Если вы отправите ему предложение по электронной почте следующим утром, будет уже слишком поздно. Вы должны сделать ему предложение, актуальное для конкретного времени и места, мимо которого он проезжает.

Делайте своевременные предложения

Новая тенденция в маркетинге — создание клиентских предложений, действующих лишь в пределах конкретного места и только в определенный период. Такие предложения могут быть гораздо более мощными и целенаправленными, чем предложения для неопределенного времени и места. Те, кто сумел применить такой подход, уже увидели эффективные результаты.

Разумеется, процесс управления предложениями усложняется, поскольку уже недостаточно просто отслеживать, какие предложения следует сделать каждому клиенту на этой неделе. Необходимо беспокоиться о том, где находится каждый клиент в любой момент и какое предложение ему следует сделать исходя из этого. Зависимыми от времени и места предложениями действительно будет труднее управлять. Однако в долгосрочной перспективе результаты, полученные при надлежащем использовании таких предложений, значительно превзойдут

результаты использования традиционных персональных предложений. История неоднократно показала, что более нацеленные и конкретные предложения получают лучший отклик.

Данные о времени и местонахождении используются и в процессе анализа социальных сетей. В дополнение к тому, что технология беспроводной связи позволяет определить взаимоотношения людей на основе голосовых или текстовых взаимодействий, данные о времени и местонахождении позволяют выявить людей, находящихся в одном и том же месте в одно и то же время. Например, кто посещал данный концерт или фильм? Кто ходил на то или иное спортивное событие? Кто обедал в конкретном ресторане в одно и то же время?

Выявив людей, которые часто оказываются в одном и том же месте в одно и то же время, можно определить тех, кто друг с другом незнаком, но принадлежит к одной и той же социальной сети и имеет много общих интересов. Представьте себе сервис знакомств, который располагает такой информацией и помогает людям найти свою судьбу! Возможно, стоит побудить людей познакомиться друг с другом или предложить им товары, которые могут быть им интересны?

Данные о времени и местонахождении помогают не только выявить прошлые закономерности, но и позволяют довольно точно предсказывать, где клиенты будут находиться в будущем. Это особенно касается тех людей, которые придерживаются определенного графика. Если вы знаете, где находится конкретный человек и куда он направляется, то на основе этой информации можете предсказать, где он окажется через 10 минут или через час. Изучив историю передвижения потребителей, вы сможете предсказать, куда они направляются, когда следуют по тому или иному маршруту. По крайней мере список возможных вариантов значительно сузится. Это поможет обеспечить лучший таргетинг.

В ближайшие годы найдутся новые способы использования данных о времени и местонахождении; будут усовершенствованы процессы предоставления разрешения на использование данных, а также стимулы для потребителей. Пока же будьте очень осторожными и постарайтесь получить у своих клиентов явное согласие на использование информации. Это сделает сообщения более целенаправленными и личными по сравнению с сегодняшними. Возможно, что в скором времени идея создания предложений, которые не относятся к текущему времени и месту, покажется устаревшей.

Розничная торговля и производство: значение данных радиочастотной идентификации (RFID)

Метка радиочастотной идентификации, или RFID-метка, — это небольшая метка, которая помещается на палеты или упаковки с товарами. RFID-метка содержит уникальный серийный номер в отличие от UPC — общего кода для идентификации того или иного товара. Другими словами, она определяет, что в данной палете находится не просто несколько компьютеров модели 123, а уникальный набор компьютеров модели 123.

Когда считывающее устройство посылает сигнал, RFID-метка в ответ отправляет информацию. Если в радиусе действия этого устройства находится несколько меток, то все они могут ответить на один и тот же запрос, что существенно облегчает учет множества предметов. Даже если предметы располагаются друг на друге или за стенами, до тех пор пока сигналы считывающего устройства достигают меток, от них можно получить ответ. RFID-метки устраняют необходимость вручную пересчитывать все предметы и позволяют намного быстрее производить инвентаризацию.

Большинство используемых RFID-меток известны как пассивные. Это означает, что такие метки не имеют встроенных источников питания. Радиоволны считывающего устройства создают магнитное поле, обеспечивающее достаточную мощность, которая позволяет метке отправить содержащуюся в ней информацию. Хотя RFID-технология существует уже довольно длительное время, ее стоимость не позволяла применять ее повсеместно. Сегодня пассивная метка стоит всего несколько центов, и цена продолжает снижаться. По мере дальнейшего снижения цен сфера применения этой технологии будет продолжать расширяться. На сегодняшний день с RFID-технологией связаны определенные технические проблемы. Одна из них заключается в том, что жидкости могут блокировать сигналы. Со временем эти вопросы должны решиться путем обновления используемых технологий.

Существуют способы использования радиочастотной идентификации, знакомые большинству людей. Один из них — автоматический сбор оплаты проезда по платным дорогам, позволяющий водителям не останавливаться, проезжая мимо пункта взимания платы. В карте, предоставленной *органом*, взимающим плату, присутствует RFID-метка. Вдоль дороги расположены считывающие устройства. Когда автомобиль

проезжает мимо, метка передает данные об автомобиле, что обеспечивает регистрацию факта вашего проезда.

Широко известен такой способ использования RFID-данных, как контроль за имуществом. Так, например, организация может пометить каждый принадлежащий ей ПК, рабочий стол или телевизор. Такие метки обеспечивают надежную инвентаризацию, а также оповещают, если предметы перемещаются за пределы предусмотренных областей. Например, считывающие устройства могут быть размещены на выходах. Если корпоративный актив покидает пределы организации без предварительного разрешения, подается сигнал, предупреждающий службу безопасности. Точно так же в магазинах метки подают сигнал тревоги, если их не деактивировали.

Один из самых популярных способов использования технологии RFID на сегодняшний день — отслеживание предметов и палет в пределах производственных и торговых площадей. Так, например, метка может быть помещена на каждую палету, отправляемую производителем розничному торговцу. Это облегчает учет запасов, находящихся в том или ином распределительном *центре* или магазине. В итоге практически каждый товар в магазине будет снабжен RFID-меткой или подобным чипом. Теперь, когда мы поняли, что собой представляют RFID-данные, рассмотрим, как их использование может улучшить современный бизнес.

Использование данных радиочастотной идентификации

RFID-данные могут сигнализировать, когда на полках розничного магазина заканчивается тот или иной товар. Если считывающее устройство постоянно опрашивает полки, чтобы определить оставшееся количество единиц товара, оно может уведомить о необходимости пополнить запасы. RFID-данные позволяют гораздо лучше отслеживать наличие товаров на полках, поскольку существует большая разница между наличием товара на складе и на полке. Может быть так, что на полке товар отсутствует, в то время как на складе находится пять упаковок.

В данном случае любой из традиционных способов анализа запасов покажет, что запасов достаточно и беспокоиться не о чем. Когда показатели продаж начинают падать, люди удивляются, почему это происходит. Если на товарах есть RFID-метка, легко установить, что на складе находятся пять единиц товара, а на полке — ни одной. Товар просто

перемещается со склада на полку. Сегодня существуют определенные проблемы с точки зрения стоимости и технологии, однако для их решения предпринимаются определенные действия.

RFID-данные могут использоваться и для оценки эффективности использования рекламных стендов. Во время проведения специальных рекламных акций товар бывает представлен в нескольких местах по всему магазину. Традиционные данные, полученные из точки продажи, покажут только то, что рекламируемый товар продан. Невозможно узнать, с какого стенда он был взят. С помощью RFID-меток можно определить, какие товары были взяты с того или иного стенда. Это позволяет оценить влияние месторасположения на эффективность рекламного стенда.

Ценность RFID-данных возрастает при их комбинировании с другими данными. Если компания занимается сбором данных о температуре в распределительном центре, то товары, которые находились там во время отключения электроэнергии или другого чрезвычайного события, можно проверить на предмет порчи. Например, во время отключения электроэнергии, продолжавшегося в течение 90 минут, температура в определенной части склада сохранялась на отметке 32 °C. Благодаря технологии RFID можно точно установить, какие именно палеты находились в этой части распределительного центра в то время, и принять соответствующие меры. Затем данные склада сопоставляются с данными о доставке. В случае вероятного повреждения товаров можно отозвать их или предупредить розничных торговцев, чтобы они проверили полученные товары.

Ценность — в объединении данных

Ценность RFID-данных, как и многих других источников больших данных, заключается не только в том, что RFID-данные могут сообщить сами по себе. Ценность заключается в той информации, которую эти данные могут предоставить в сочетании с другими данными. Нелишним будет еще раз подчеркнуть, что большие данные должны быть интегрированы в те же процессы, что и другие данные. Не следует работать с большими данными отдельно от всего остального.

Существуют и оперативные способы применения RFID-данных. Сотрудники распределительных центров не всегда аккуратно обращаются с товарами, и это часто приводит к их порче. Возможно, это касается конкретной бригады или даже конкретных сотрудников. Отдел по работе с персоналом сообщит о том, кто работает в тот или иной

момент. Объединив эти данные с RFID-данными, которые показывают, когда товар был перемещен, можно определить тех сотрудников, чья работа сопровождается необычно частыми случаями поломки, усадки и воровства. Сочетание данных позволяет предпринимать более эффективные действия.

В будущем появится возможность использовать RFID-данные для отслеживания процесса покупки в физическом магазине так же, как это делается в интернет-магазинах. Если считывающие устройства будут находиться в тележках, можно определить, какие товары и в каком порядке потребители помещают в свои тележки. Даже если на отдельных товарах отсутствует метка, нетрудно установить путь перемещения корзины. При таком способе применения RFID-данных в физическом магазине становится возможным использование многих преимуществ, которые предоставляют веб-данные, о чем шла речь в главе 2. Эти последние два примера снова затрагивают проблемы конфиденциальности. Возможно, потребители не захотят, чтобы их процесс покупки отслеживался. В этом случае можно наладить «анонимное» отслеживание процессов покупки, при котором личность потребителя, генерирующего данные, нельзя установить.

Последний способ применения технологии RFID касается того, как можно уменьшить количество случаев мошенничества, связанных с возвратом украденных вещей. Если товар имеет RFID-метку, то благодаря уникальному идентификатору метки магазин определит, что возвращаемый товар входил в украденную партию, и примет соответствующие меры. Со временем RFID-метка может стать частью чека и ее будут запрашивать при возврате товара. Магазины будут известно не только то, что вы купили определенный товар, но и какая конкретно RFID-метка на нем находилась. Когда вы придете в пункт возврата товара, вам предстоит вернуть конкретный товар с конкретной меткой. Вы не сможете взять другой такой же товар с полки и обманным путем вернуть его вместе со своим чеком. Подобный способ использования технологии RFID затруднит попытки совершения мошеннических действий.

В ближайшие годы технология RFID окажет огромное влияние на производство и розничную торговлю. Она завоевывала популярность медленнее, чем многие ожидали. Однако по мере снижения стоимости и улучшения качества меток и считывающих устройств эта технология получит более широкое распространение.

Коммунальные предприятия: значение данных, генерируемых интеллектуальными сетями

Интеллектуальные сети представляют собой новое поколение электро-энергетической инфраструктуры. Интеллектуальная сеть гораздо более совершенна и надежна, чем традиционные линии электропередач. Она предполагает наличие сложных систем мониторинга, связи и генерации энергии, которые обеспечивают более надежное обслуживание и восстановление после отключения питания или возникновения других проблем. Различные датчики и мониторы отслеживают множество показателей работы самой энергосистемы и электроэнергии, подающейся через нее.

Одно из нововведений — так называемый интеллектуальный счетчик, который сменил традиционные электрические счетчики. По виду интеллектуальный счетчик мало чем отличается от привычных, однако он гораздо более функционален. Прежде каждые несколько недель или месяцев необходимо было посещать то или иное помещение и фиксировать показатели потребления электроэнергии. Интеллектуальный счетчик автоматически собирает данные, как правило, каждые 15 минут или каждый час. Это позволяет более точно оценивать энергопотребление каждой семьи или предприятия, а также целого района или даже всей сети.

Хотя разговор будет сосредоточен на интеллектуальных счетчиках, следует упомянуть о датчиках, расположенных по всей интеллектуальной сети. Объем данных, которые коммунальные предприятия получают от датчиков, установленных по всей интеллектуальной сети, значительно превышает объем данных, получаемых от интеллектуальных счетчиков. Синхрофазоры, которые снимают 60 показаний о работе энергетической системы в секунду, а также домашние сети, фиксирующие работу каждого устройства, — лишь два примера. Средний потребитель и не догадывается о существовании таких датчиков, однако для предприятий коммунального обслуживания они возымеют решающее значение. Эти датчики смогут собирать полный спектр данных о состоянии всей энергосистемы. Объем этих данных будет огромным.

Интеллектуальные сети уже применяются в некоторых странах Европы и Америки. Со временем практически каждую электрическую сеть в мире заменит интеллектуальная. Объем данных о потреблении

электроэнергии, которые в результате станут доступными для коммунальных предприятий, вырастет в геометрической прогрессии. Как можно использовать такие данные? Давайте разберемся.

Использование данных интеллектуальных сетей

С точки зрения управления питанием данные от интеллектуальных счетчиков помогут лучше понять уровень потребностей клиентов, а также предоставить некоторые преимущества потребителям. Любой домовладелец сможет, например, проверить, какую мощность потребляют различные приборы, включив их по очереди, а затем изучив статистику потребления электроэнергии, предоставляемую интеллектуальным счетчиком.

Предприятия коммунального обслуживания по всему миру уже активно переходят на использование моделей ценообразования, учитывающих время суток или уровень спроса, а распространение интеллектуальных сетей ускорит этот процесс. Одна из основных задач коммунальных предприятий заключается в использовании новых программ ценообразования с целью повлиять на поведение клиентов и сократить потребление в часы максимальной нагрузки. Именно эти периоды пиковой нагрузки заставляют предприятия наращивать генерируемые мощности, что требует существенных затрат и оказывает негативное воздействие на окружающую среду. Если стоимость электроэнергии будет меняться в зависимости от времени суток и измеряться счетчиком, то у потребителей появится стимул изменить свое поведение. Снижение нагрузки и поддержание более стабильного уровня потребления приведут к уменьшению необходимости в расширении инфраструктуры и снижению затрат.

Энергетическая компания сможет выявлять всевозможные дополнительные тенденции, анализируя данные, предоставляемые интеллектуальными счетчиками. Какие клиенты потребляют мощность в периоды относительного спада уровня потребления? Кто из клиентов имеет похожие энергетические потребности в течение дня или недели? Коммунальное предприятие получит возможность сегментировать клиентов, исходя из закономерностей потребления, и разрабатывать продукты и программы для каждого конкретного сегмента. Собранные данные позволят выявить необычные модели потребления, а они укажут, какие проблемы требуют коррекции.

По сути, энергетические компании смогут производить анализ клиентов, который уже давно делается во всех остальных отраслях. Представьте себе телефонную компанию, которой известна общая сумма по счету в конце месяца, но ничего не известно о сделанных вами звонках. Или розничный магазин, который знает только общую сумму ваших покупок — и никаких дополнительных сведений. Финансовое учреждение, которому известен баланс по вашему счету на конец месяца — и никаких деталей о движении средств в течение месяца. Коммунальные компании работали с подобными данными, которых было недостаточно для понимания поведения клиентов. Они располагали данными об общем потреблении за месяц, и даже этот показатель часто был предположительным, а не точным.

Большие данные могут кардинально изменить отрасль

В некоторых случаях большие данные в буквальном смысле трансформируют отрасль и позволяют вывести аналитику на новый уровень. Данные интеллектуальных сетей в сфере коммунального обслуживания — это только один из примеров. Информация об использовании электроэнергии будет поступать не раз в месяц, а предоставляться с интервалом, измеряемым в секундах или минутах. Добавьте к этому наличие сложных датчиков по всей сети, и вы увидите совершенно другой мир с точки зрения данных. Анализ этих данных приведет к инновациям в тарифных планах, управлении питанием, и не только.

Данные, полученные от интеллектуальных счетчиков, позволят применить целый ряд новых способов анализа, от чего выиграют все. Потребители получают возможность использовать индивидуальные тарифные планы, основанные на их индивидуальных закономерностях потребления, подобно тому как телематические данные позволяют использовать индивидуальные тарифы автострахования. Клиент, который потребляет электроэнергию во время периодов пиковой нагрузки, будет платить больше, чем другие. Это заставит всех нас изменить модели потребления, как только мы получим соответствующие стимулы: например, мы будем включать посудомоечную машину в конце дня, а не сразу после обеда.

Коммунальные предприятия смогут лучше прогнозировать спрос, поскольку будут более подробно определять его закономерности. Они будут знать, какие именно клиенты потребляют мощность в тот или иной момент. Поставщик коммунальных услуг найдет способы влияния на поведение клиентов, чтобы выровнять уровень спроса и снизить

частоту возникновения периодов пиковой нагрузки. Все это уменьшит необходимость в наращивании дорогостоящих генерирующих мощностей.

Благодаря интеллектуальным счетчикам каждое домохозяйство или предприятие получит возможность более эффективно отслеживать свое энергопотребление и принимать соответствующие меры. Это позволит не только более рационально использовать энергию и беречь окружающую среду, но и экономить деньги. В конце концов, если вы будете в состоянии определить, что потратили больше, чем собирались, то сможете скорректировать свое дальнейшее поведение. Наличие одного только итогового счета в конце месяца не позволит выявить такие возможности. Данные интеллектуального счетчика облегчают решение задачи.

Индустрия игр: значение данных отслеживания фишек

Мы уже рассказали о технологии радиочастотной идентификации и о том, как она используется в розничной торговле и производстве. Однако технология RFID имеет широкий спектр способов применения, и многие из них также приводят к генерации больших данных. В частности, RFID-метки могут использоваться в фишках казино. Каждая фишка, особенно высокой стоимости, может иметь собственную встроенную метку, что позволяет однозначно идентифицировать ее по серийному номеру метки.

Работа игровых автоматов в казино отслеживается уже в течение многих лет. После того как вы вставите в автомат свою карту постоянного игрока или кредитную карту, а затем потянете за ручку или нажмете кнопку, этот факт будет зарегистрирован. Кроме того, регистрируются размеры ваших ставок, а также любые полученные вами выигрыши. На протяжении многих лет производился надежный анализ таких данных, однако не существовало возможности сбора подобных данных с игровых столов. Внедрение меток в игровые фишки позволит это сделать.

Казино всегда отслеживало фишки с помощью камер и сотрудников, которые находились на местах и контролировали их перемещение.

Пит-босс* наблюдает за часто играющими игроками и оценивает их средние ставки и продолжительность игры, чтобы обеспечить соответствующие поощрения. Пит-боссы хорошо справляются со своей задачей, к тому же им помогают другие сотрудники, тем не менее они могут переоценить или недооценить игрока. Это происходит, если пит-босс наблюдает за игроком в тот момент, когда он ставит больше или меньше, чем обычно. Некоторые игроки даже пытаются обыграть систему, увеличивая свои ставки тогда, когда за ними, по их мнению, наблюдают.

Одна и та же технология может генерировать множество потоков больших данных

Розничные торговцы и производители используют технологию радиочастотной идентификации, как и владельцы казино. Способы ее применения имеют сходства и различия. Интереснее всего то, что одна и та же технология может быть использована в различных отраслях для создания различных источников больших данных.

Пример с фишками казино — уникальный, но не исключительный способ применения технологии RFID. Будут появляться и другие. Этот пример показывает, что одни и те же базовые технологии генерируют различные потоки больших данных, схожих по природе, но различающихся способом применения. Интересно, что одна и та же технология может иметь совершенно разные виды применения, генерирующие множество типов больших данных в различных отраслях.

Использование данных отслеживания фишек казино

Использование меток в фишках казино дает возможность точно отслеживать ставки каждого игрока, что гарантирует игроку заслуженный выигрыш в соответствии с условиями программы, награждающей постоянных игроков. От этого получают выгоду и казино, и игрок. Казино распределяет средства среди игроков более справедливо. Излишнее поощрение недостойных игроков и недостаточное поощрение достойных приводят к неоптимальному распределению ограниченных ресурсов,

* Пит-босс — сотрудник казино, контролирующий порядок и соблюдение правил честной игры в рамках пита — нескольких игровых столов. *Прим. ред.*

выделенных на маркетинговые программы. Игроки, разумеется, всегда хотят, чтобы их оценивали справедливо.

Данные о ставках игроков позволят казино лучше их сегментировать и понимать закономерности назначения ставок. Кто из игроков обычно ставит \$5, но время от времени поднимает ставку до \$100? Кто из игроков каждый раз ставит по \$10? Игроков можно сегментировать исходя из этих закономерностей. Кроме того, закономерности в назначении ставок могут указать на тех, кто занимается подсчетом карт при игре в блэк-джек.

При отслеживании фишек игроку становится намного тяжелее намеренно обмануть казино, а крупье — допустить ошибку. Поскольку ставки и выплаты можно проследить по фишкам, легко вернуться и сравнить видеозапись с результатами сдачи карт при игре в блэк-джек и сделанные выплаты. Даже если руки и головы игроков не позволяют разглядеть фишки, RFID-данные предоставят подробную информацию, и казино сможет выявить ошибки или случаи мошенничества, которые имели место. Один из примеров — ситуация, когда игрок подкладывает дополнительные фишки, пока крупье этого не видит.

Анализ за определенный период позволит выявить крупье или игроков, допускающих необычно большое количество ошибок. После этого либо будут приняты меры по предотвращению мошенничества, либо сотрудников обяжут пройти дополнительное обучение. Кроме того, снизится количество ошибок при подсчете фишек в кассе казино. Подсчет большого количества фишек разного достоинства — весьма однообразное занятие, и люди допускают ошибки. Технология RFID обеспечит более быстрый и точный подсчет.

Следует отметить, что возможность отслеживания отдельных фишек окажется сдерживающим фактором для воров. При краже стопки фишек идентификаторы RFID сообщат об этом. Когда кто-то приходит, чтобы обналечить фишки, или просто сидит с ними за столом, система выявляет этот факт и предупреждает службу безопасности. Если воры изменяют фишки так, что информацию с них невозможно было считать, об этом тоже станет известно. Казино будет точно знать все существующие идентификаторы фишек и ожидать от каждой из них сообщения действительного ID. В случае когда фишка не сообщает свой ID или когда сообщенный ID недействителен, казино может принять соответствующие меры.

Как и в любом бизнесе, чем эффективнее казино справляется с мошенничеством и обеспечивает соответствующие выплаты, тем меньше

у него рисков. Это означает лучшее обслуживание игроков и более высокие шансы на выигрыш, поскольку у казино будет меньше расходов. Это выгодно и казино, и игрокам.

Промышленные двигатели и оборудование: значение данных, полученных от датчиков

В мире существует множество сложных машин и двигателей: самолеты, поезда, военная техника, строительное оборудование, буровое оборудование и т. д. Обеспечение надежной работы этого оборудования имеет огромное значение. В последние годы во всех видах техники — от двигателей самолетов до танков — начали использоваться встроенные датчики с целью посекундного наблюдения за состоянием оборудования.

Мониторинг может осуществляться с огромной степенью детализации, особенно во время тестирования и разработки. Например, в процессе разработки нового двигателя следует собрать как можно более подробные данные, чтобы определить, работает ли он так, как ожидалось. После выпуска двигателя замена в нем того или иного компонента связана с большими затратами, поэтому необходимо заранее тщательно проанализировать работу оборудования. Мониторинг также никогда не прекращается. Возможно, при постоянном мониторинге собираются не все данные, поступающие каждую миллисекунду, однако большое количество детальных сведений собирается для оценки жизненного цикла оборудования и выявления регулярно возникающих проблем.

Возьмем двигатель. Датчик может собирать все данные, начиная от температуры и количества оборотов в минуту и заканчивая скоростью расхода топлива и уровнем давления масла, с необходимой частотой. Объем данных очень быстро растет по мере увеличения частоты получения данных, количества метрик и числа отслеживаемых элементов. Почему это должно нас заинтересовать? Вот несколько примеров.

Использование данных, полученных от датчиков

Двигатели — очень сложные устройства. Они содержат много движущихся частей, должны работать при высоких температурах и в широком диапазоне рабочих условий. Их стоимость предполагает, что они проработают много лет. Стабильная и предсказуемая работа имеет решающее

значение, и часто от нее зависят жизни людей. Техническое обслуживание самолета может стоить авиакомпании или военно-воздушным силам страны огромных денег, но это необходимо, если под угрозой безопасность. Крайне важно минимизировать время, в течение которого самолеты и авиационные двигатели, а также другое оборудование оказываются выведенными из эксплуатации.

Стратегии минимизации времени простоя предусматривают наличие запчастей или дополнительных двигателей, которые можно быстро заменить в оборудовании, требующем обслуживания; проведение диагностики для быстрого обнаружения частей, требующих замены, и закупку более надежных версий проблемных компонентов. Эффективное применение всех этих стратегий зависит от данных, которые используются для создания диагностических алгоритмов, а также в качестве входных данных при проведении диагностики определенных проблем. Инженерные организации с помощью данных, полученных от датчиков, точно определяют причины выхода оборудования из строя и разрабатывают новые способы обеспечения более длительной и надежной работы. Эти соображения применимы к двигателям воздушного, водного и наземного транспорта.

Путем сбора и анализа подробных данных о работе двигателя можно точно определить закономерности, которые приводят к поломке. Кроме того, можно выявить долгосрочные закономерности, которые вызывают уменьшение срока службы двигателя и/или необходимость более частого проведения ремонта. Количество пермутаций различных показаний, особенно с течением времени, делает анализ этих данных настоящим вызовом. Мало того что этот процесс подразумевает работу с большими данными, но и анализ, который необходимо проводить, сам по себе очень сложен. Вот некоторые примеры вопросов, на которые можно найти ответ:

- Предсказывает ли внезапное падение давления неизбежный отказ оборудования почти со 100%-ной вероятностью?
- Указывает ли устойчивое снижение температуры в течение нескольких часов на другие проблемы?
- Что означают необычные уровни вибрации?
- Приводит ли резкое увеличение оборотов двигателя при запуске к значительному изнашиванию определенных компонентов и увеличению частоты проведения требуемого ремонта?

- ▶ Приводит ли недостаточное давление топлива, сохраняющееся на протяжении нескольких месяцев, к повреждению некоторых компонентов двигателя?

Недостаток структуры в структурированных данных

Работа с информацией, полученной от датчиков, — непростая задача. Хотя собранные данные структурированы и отдельные их элементы достаточно изучены, взаимосвязи между этими элементами не очевидны. Задержки во времени и не поддающиеся измерению внешние факторы могут еще больше усложнить дело. Процесс выявления долгосрочных взаимодействий различных показателей чрезвычайно труден, учитывая объем доступной информации. Наличие структурированных данных не гарантирует высокоструктурированного и стандартизированного подхода к их анализу.

Когда возникают серьезные проблемы, очень полезно вернуться и посмотреть, что происходило перед тем, как проблема была выявлена. В данном случае датчики двигателя работают подобно черным ящикам самолета, которые помогают диагностировать причину аварии: полученные данные можно использовать в диагностических и исследовательских целях. Такие датчики представляют собой более сложную форму телематических устройств — о них упоминалось в примере с автострахованием. Использование данных от сенсоров, которые постоянно собирают информацию о своем окружении, широко обсуждается в мире больших данных. Хотя здесь мы сосредоточили внимание на двигателях, существует бесчисленное множество других способов использования датчиков, к которым применимы те же принципы.

Процесс сбора данных от датчиков множества двигателей в течение длительного периода обеспечивает большой объем данных для анализа. Тщательно изучив их, вы сможете выявить проблемы в оборудовании и решить их на ранней стадии. Можно определить слабые стороны, а затем разработать процедуры решения проблем, которые могут возникнуть в результате выявленных недостатков. Преимущества заключаются не только в повышении уровня безопасности, но и в снижении затрат. Поскольку данные, полученные от датчиков, позволяют повысить безопасность двигателей и оборудования, которые дольше остаются в эксплуатации, это позволит обеспечить стабильную работу и снизить затраты. В этом случае выигрывают все.

Видеоигры: значение телеметрических данных

Телеметрия — это термин, используемый в индустрии видеоигр для фиксации действий в игре. Телеметрические данные имеют концептуальное сходство с веб-данными, о которых шла речь в главе 2, поскольку описывают действия игроков в процессе игры. Телеметрические данные чаще применяются в онлайн-играх, чем в игровых консолях.

При игре в хоккей телеметрические данные фиксируют такие показатели, как местоположение игрока в момент удара по воротам, тип этого удара, скорость удара и его результат. В военной игре телеметрические данные отмечают, из какого оружия и в каком направлении был произведен выстрел, какой ущерб нанесен различным объектам. Теоретически можно обеспечить любой уровень детализации данных о сцене и действии.

Это позволит определять не только количество игроков и время, в течение которого они играли в игру. Телеметрические данные дают производителям игр возможность узнать подробности о действиях игроков и о том, как они взаимодействуют с играми. Объем собранных данных может быть огромным, а индустрия видеоигр только начинает серьезно подходить к их анализу. Телеметрия может оказать влияние на множество сфер. Легко увидеть параллель между телеметрическими данными и веб-данными с точки зрения преимуществ и способов применения.

Использование телеметрических данных

Многие производители игр зарабатывают деньги с помощью подписки, поэтому для них решающее значение имеет ее обновление. Анализ моделей поведения игроков позволяет понять, какие типы игрового поведения приводят к продлению подписки, а какие нет. Например, выясняется, что проведение турниров по спортивной игре с одновременным использованием некоторых дополнительных функций увеличивает показатель продления подписки. Производитель игры может стимулировать игроков принять участие в турнире, используя эти функции, если они еще этого не сделали.

Более новые игры часто предлагают игрокам купить что-нибудь за небольшую плату. Такие покупки известны как микротранзакции.

Например, за 10 центов игрок может купить специальное оружие. Анализ данных о ходе игры позволяет определить особые области, где такие микротранзакции будут пользоваться успехом. Возможно, в определенном месте игры пригодится специальное оружие, поскольку многие игроки испытывают в этом месте трудности. Сообщение на экране о доступности оружия может привести к тому, что многие игроки примут предложение и совершат покупку.

Объем телеметрических данных будет увеличиваться

В настоящее время телеметрические данные в основном касаются действий, произведенных игроком с помощью контроллера или клавиатуры. По мере развития интерактивных игр, которые позволяют фиксировать движения самого игрока, а не контроллеров, объем данных будет стремительно увеличиваться. Данные о том, на какую кнопку нажал игрок в тот или иной момент, предоставляют гораздо меньше информации, чем данные о том, в какой точке пространства находилась в этот момент каждая из частей тела игрока и в каком направлении и с какой скоростью эта часть тела двигалась.

Удовлетворение потребностей клиентов в индустрии видеоигр столь же важно, как и в любой другой сфере. Вот только в этом случае грань очень тонкая. Игра должна ставить перед игроками сложную задачу, однако не настолько сложную, чтобы она надоела игрокам и они выбрали бы другую.

Путем анализа игры можно определить те части, которые легко преодолевают большинство игроков, а также части, в которых даже лучшие игроки испытывают трудности. Такие области можно скорректировать, например, увеличив или уменьшив количество врагов, чтобы изменить уровень сложности. Стабилизация уровня сложности игры обеспечивает пользователям более удовлетворительный игровой опыт. Это приведет к повышению показателей продления подписки и к увеличению объема дополнительных покупок.

С помощью телеметрических данных можно сегментировать игроков исходя из их стиля игры. Эта информация важна как для разработки новых игр, так и для продвижения других существующих продуктов. Например, выясняется, что один сегмент игроков старается как можно быстрее преодолеть уровень, не заботясь ни о чем другом. Другой пытается собрать все бонусы перед завершением уровня. Третий сегмент стремится исследовать каждую деталь уровня. На основе этих данных можно осведомить игроков о других играх, соответствующих их стилю игры.

Сведения об игроках, которые могут предоставить телеметрические данные, полностью изменят индустрию видеоигр, которая только начинает использовать телеметрические данные. Уже в ближайшем будущем мы станем свидетелями значительного развития данной области применения. Кроме того, полученные в результате анализа телеметрических данных сведения изменяют процесс создания и продвижения игр.

Телекоммуникации и другие отрасли: значение данных, полученных из социальных сетей

Социальные сети — источники больших данных, хотя во многих отношениях речь идет скорее о методологии анализа традиционных данных. Дело в том, что процесс анализа социальных сетей подразумевает работу с очень большими наборами данных и их использование таким способом, который увеличивает этот объем на несколько порядков.

Можно утверждать, что полный набор звонков по мобильному телефону или история текстовых сообщений, собранных оператором мобильной связи, сами по себе представляют источник больших данных. Анализ социальных сетей выведет их использование на новый уровень путем изучения нескольких видов ассоциаций вместо одного. Именно поэтому анализ социальных сетей может превратить источники традиционных данных в источники больших данных.

Современной телефонной компании уже недостаточно просто анализировать все звонки по отдельности. При анализе социальной сети необходимо определить, кто был участником телефонного разговора, а затем провести более глубокое изучение. Нужно узнать не только кому звонил я, но и кому, в свою очередь, звонили эти люди, кому звонили те люди и т. д. Чтобы получить более полное представление о социальной сети, можно проанализировать столько слоев, сколько позволяет система. При переходе от клиента к клиенту и от звонка к звонку объем данных возрастает в несколько раз. Это также усложняет их анализ, особенно когда речь идет о традиционных инструментах.

Те же принципы применимы к сайтам социальных сетей. При анализе любого пользователя социальной сети нетрудно определить, сколько у него контактов, как часто он отправляет сообщения, как часто заходит на сайт, а также другие стандартные метрики. Однако

анализ широты сети контактов данного участника, включая его друзей, друзей друзей и друзей друзей друзей, предполагает гораздо более сложную обработку.

Нетрудно отследить одну тысячу участников или подписчиков. Однако между ними могут существовать до миллиона прямых и до миллиарда косвенных связей, если учитывать «друзей друзей». Именно поэтому анализ социальных сетей подразумевает работу с большими данными. На сегодняшний день он имеет целый ряд приложений.

Использование данных социальной сети

Данные социальных сетей и их анализ могут быть очень полезны, например, для изменения взгляда организации на своих клиентов. Отныне во главу угла будет ставиться не платежеспособность отдельного потребителя, а ценность его сети контактов. Пример, о котором пойдет речь, применим во многих других отраслях, где известны отношения между людьми или группами, однако мы сосредоточимся на беспроводных телефонах, поскольку именно в этой сфере данные методы используются наиболее широко.

Предположим, у оператора беспроводной связи есть абонент с относительно низкой платежеспособностью. Он пользуется базовым тарифным планом и не прибегает к дополнительным услугам. Этого клиента едва ли можно считать прибыльным. Традиционно оператор стал бы оценивать его на основании индивидуального счета, и, если бы такой клиент позвонил, чтобы пожаловаться, и стал угрожать расторжением договора на обслуживание, компания, скорее всего, позволила бы ему уйти, поскольку данный потребитель просто не стоит того, чтобы его удерживать.

Анализ социальной сети может выявить тот факт, что среди тех, кому звонит наш клиент, есть очень активные пользователи, имеющие весьма широкий круг друзей. Другими словами, контакты данного клиента имеют очень большую ценность для организации. Исследования показали, что, если один человек из этого круга покидает его, другие, скорее всего, последуют за ним. Выход людей из этого круга может приобрести характер эпидемии, и вскоре количество его участников будет стремительно сокращаться, что, безусловно, крайне нежелательно.

Учитывайте не только индивидуальную ценность

Очень важное преимущество использования данных социальной сети заключается в том, что она предоставляет возможность определить общий доход, на который влияет клиент, а не только прямой доход, который этот клиент генерирует. Это может привести к принятию совершенно иных решений о том, как следует инвестировать в этого клиента. Клиента, который обладает большим влиянием, необходимо поощрять гораздо больше, если, конечно, максимизация общей прибыли для организации важнее, чем максимизация прибыльности отдельных клиентов.

С помощью анализа социальной сети можно оценить общий доход организации, на который влияет клиент в нашем примере, а не только выручку, непосредственно им генерируемую. Это позволяет принять совершенно иные решения о том, как следует обращаться с этим клиентом. Оператор беспроводной связи может вложить в этого клиента дополнительные средства, чтобы защитить сеть, в которую он входит. Можно обеспечить стимулы, превосходящие индивидуальную доходность клиента, если это позволит заинтересовать более широкий круг клиентов, в который он входит.

Это замечательный пример того, что благодаря анализу больших данных приходят решения, в прошлом немислимые, и обретают смысл. Без больших данных организация не попыталась бы удержать этого клиента и не осознала бы причину убытков, которые вскоре начали бы проявляться по мере того, как друзья этого клиента следовали бы его примеру. Цель сдвигается от максимизации прибыльности отдельных счетов к максимизации прибыльности сети потребителя.

Выявление клиентов, обладающих большой сетью контактов, позволяет определить, где следует сосредоточить усилия по укреплению имиджа бренда. Потребителям с большими связями можно предоставить бесплатные пробные версии продуктов в обмен на их отзывы. Стимулы помогут привлечь их к активному участию в обсуждениях на сайте корпоративной социальной сети, где они могли бы оставлять комментарии и мнения. Некоторые организации активно вербуют влиятельных клиентов и предоставляют им различные льготы, возможность раньше других испытать пробные версии продуктов и т. д. В свою очередь, эти клиенты продолжают оказывать влияние, тон которого становится все более позитивным, учитывая особое отношение к ним со стороны компании.

Анализ, проведенный в рамках таких социальных сетей, как LinkedIn или Facebook, помогает понять, с какой рекламой имеет смысл обра-

щаться к конкретным пользователям. При этом учитываются не только те интересы, о которых они заявили лично. Не менее важны интересы их круга друзей или коллег. Пользователи никогда не сообщают обо всех своих интересах в социальных сетях, и невозможно узнать о них все подробности. Тем не менее, если большинство друзей пользователя интересуются, к примеру, велосипедным спортом, весьма вероятно, что и данный пользователь им тоже интересуется, даже если он не заявляет об этом прямо.

Анализ социальных сетей может быть полезен в целях борьбы с преступностью и предотвращения террористических актов. Можно выявить людей, связанных, пусть даже косвенно, с известными проблемными группами или лицами. Такой анализ называется анализом связей. Проблемными могут быть как физические лица и группы людей, так и клуб или ресторан. Если в результате анализа выявляется, что данный человек часто общается с данными людьми в данных местах, то к нему следует присмотреться более внимательно. Хотя подобный вид анализа связан с проблемами конфиденциальности, он применяется сегодня в реальных жизненных ситуациях.

Этот вид анализа может оказаться полезным в онлайн-видеоиграх. Кто с кем играет? Как эта закономерность изменяется от игры к игре? Анализ социальных сетей дополнит телеметрические данные. Мы определим модели, используемые игроком в каждой игре. Мы уже говорили о том, как игроков можно сегментировать исходя из индивидуального стиля игры. Объединяются ли игроки, использующие похожие игровые стили, в команды, когда играют вместе? Или они стремятся обеспечить разнообразие стилей? Такие сведения очень ценны для производителя игр, если он намерен предложить игрокам группы, к которым они могут присоединиться (например, при входе в систему пользователю предлагают конкретную группу из множества доступных вариантов).

В организациях был проведен ряд интересных исследований связей. Они начинались с изучения контактов, установленных с помощью электронной почты, телефона и текстовых сообщений в рамках организации. Взаимодействуют ли отделы друг с другом так, как ожидалось? Приходится ли некоторым сотрудникам выходить за пределы типичных каналов, чтобы решить рабочий вопрос? Кто пользуется большим влиянием и подходит для участия в исследовании способов улучшения системы коммуникации в рамках организации? Такой анализ поможет

организациям лучше понять, как взаимодействуют между собой их сотрудники.

Сферы применения и влияния анализа социальных сетей будут только расширяться. Такой анализ всегда способствует значительному увеличению объема данных благодаря экспоненциально расширяющемуся характеру аналитического процесса. Вероятно, самая интересная особенность этого вида анализа состоит в том, что он помогает оценить влияние потребителя и его общую ценность для организации, что может полностью изменить отношение со стороны организации.

Обзор главы

Самые важные уроки этой главы.

- Несмотря на существование широкого спектра источников больших данных в различных отраслях, между ними есть определенные сходства. Одни и те же базовые технологии, такие как RFID, могут быть использованы в различных отраслях для решения различных задач.
- Со многими источниками больших данных связаны проблемы соблюдения конфиденциальности, которым всегда следует уделять серьезное внимание.
- Телематические данные могут использоваться в процессе определения тарифов при страховании автомобилей. Собранные данные также могут коренным образом изменить систему планирования и управления дорожным движением.
- Текстовые данные — один из самых мощных и широко применяемых типов больших данных. В центре внимания, как правило, находится извлечение из текста ключевых фактов, которые затем используются в качестве входных данных в других аналитических процессах.
- Данные о времени и местонахождении завоевывают все большую популярность. Организациям предстоит постараться, чтобы обеспечивать потребителей предложениями, актуальными для конкретного времени и места.
- RFID-данные предоставляют розничным торговцам и производителям новые возможности проведения анализа в сферах

управления запасами, предотвращения случаев мошенничества и оценки работы сотрудников.

- ▶ Интеллектуальные сети обеспечивают возможность коммунальным предприятиям гораздо лучше управлять их энергосистемами, а потребителям — лучше контролировать свое потребление.
- ▶ Фишки с RFID-метками помогают точнее отслеживать активность игроков казино и предотвращать случаи мошенничества и ошибочные выплаты.
- ▶ Данные, полученные от датчиков, предоставляют подробную информацию о работе двигателей и оборудования. Это позволяет точнее диагностировать проблемы и быстрее принимать соответствующие меры.
- ▶ Телеметрические данные позволят производителям видеоигр лучше управлять микротранзакциями, вносить в игру улучшения и сегментировать игроков исходя из их игрового стиля.
- ▶ Данные социальных сетей могут обеспечить новые способы оценки потребителей. В телекоммуникационной отрасли фокус анализа социальных сетей сместился с оценки прибыльности счета к оценке прибыльности сети.

ЧАСТЬ II

Укрощение больших данных: технологии, процессы и методы

Эволюция масштабируемости аналитических систем

Само собой разумеется, что мир больших данных требует иных уровней масштабируемости. По мере увеличения объема данных, которые необходимо обрабатывать организациям, старые методы перестают работать. Организации, не обновляющие свои технологии, чтобы обеспечить более высокий уровень масштабируемости, просто не справятся с большими данными. К счастью, существует множество технологий, созданных для укрощения больших данных и их использования в аналитических процессах. Некоторые из них совершенно новые, и организациям следует идти в ногу со временем.

В этой главе мы расскажем о технологиях, которые позволяют добиться успеха в процессе приручения больших данных. Поговорим о слиянии аналитической среды со средой данных, массивно-параллельных архитектурах (МРР), облачных и грид-вычислениях, а также о модели MapReduce.

Хочу еще раз напомнить, что данная книга — не практическое пособие. Эта глава, а также главы 5 и 6 посвящены техническим деталям, однако они рассматриваются на концептуальном уровне, поэтому вне зависимости от подготовки вы сможете разобраться в концепциях. Некоторые вопросы излагаются очень упрощенно. При необходимости более детально разобраться в той или иной теме вы можете обратиться к другим, специализированным книгам.

История масштабируемости

Вплоть до 1900-х годов проведение анализа было очень, очень трудной задачей. Глубокий анализ, связанный, скажем, с моделью прогнозирования, требовал вычисления вручную всех статистических данных. Для выполнения линейной регрессии следовало вручную вычислить матрицу, а затем — также вручную — обратную матрицу. Все остальные вычисления, необходимые для оценки параметров, производились вручную или с помощью простейших механических счетных машин. Сто и более лет назад собрать данные было очень трудно, но использовать их было еще труднее. Ни о какой масштабируемости не могло быть и речи.

Логарифмические линейки со временем облегчили положение, а в 1970-х годах появились калькуляторы, которые упростили процесс использования большего объема данных. Однако объем данных, который можно обработать с помощью калькулятора, все еще был очень небольшим. Компьютеры, которые получили распространение в 1980-х годах, позволили людям навсегда забыть о производстве вычислений вручную. Хотя компьютеры существовали и до 1980-х годов, они были доступны только избранным, а работа с ними была сложна и связана с большими затратами.

По прошествии десятилетий объем данных вышел далеко за рамки того, что человек в состоянии обработать вручную. Количество данных росло так же быстро, как и вычислительная мощность машин, которые использовались для их обработки. И хотя нет необходимости напрягаться и до головной боли производить вычисления вручную, очень легко перегрузить компьютер и систему хранения данных.

Технология продолжала развиваться такими темпами, что объем данных возрастал год от года. Десять лет назад обработка терабайта данных была доступна только крупнейшим компаниям и богатейшим правительствам. В 2000 году организация, у которой в базе данных содержался терабайт данных, считалась передовой. Теперь вы можете купить терабайтный диск для вашего компьютера за \$100! В 2012 году даже многие небольшие компании имели системы, содержащие терабайт данных или больше. Объем данных передовых компаний сегодня измеряется в петабайтах*. Это тысячекратное увеличение всего за десять лет!

* Петабайт — единица измерения количества информации, равная 10^{15} байт.

Более того, появление новых источников больших данных раздвигает границы еще шире. Многие источники больших данных могут генерировать от терабайта до петабайта данных за несколько дней или недель, а то и часов. В настоящее время испытываются пределы того, что может быть обработано. Тем не менее со временем волна больших данных будет укрощена так же, как в прошлом были укрощены другие пугающие потоки данных.

Когда сегодняшние первоклассники пойдут в колледж, у них, вероятно, будет петабайтный компьютер и они будут работать с системами хранения данных, вмещающими эксабайты*, если не зеттабайты** информации. Они также будут ожидать получения ответов в течение секунд или минут, а не часов или дней. В табл. 4.1 приведены известные на сегодняшний день единицы измерения объема данных, а также значения, следующие за ними по данной шкале. Люди, которые первыми использовали предельные объемы данных, получали в прошлом значительные преимущества, и так будет продолжаться и в дальнейшем.

Табл. 4.1

Единицы измерения объема данных

1024...	...соответствует 1	Комментарий ⁶
Килобайт	Мегабайт	Стандартный музыкальный компакт-диск вмещает 600 мегабайт
Мегабайт	Гигабайт	Один гигабайт вмещает объем данных, эквивалентный объему информации, содержащейся в книгах, выставленных на полке длиной примерно 9 метров
Гигабайт	Терабайт	Десять терабайт могут вместить всю Библиотеку Конгресса США
Терабайт	Петабайт	Один петабайт может вместить 20 миллионов 4-дверных картотечных шкафов текста
Петабайт	Эксабайт	Пять эксабайт соответствует всем словам, когда-либо произнесенным человечеством
Эксабайт	Зеттабайт	Загрузка из интернета файла объемом в 1 зеттабайт при использовании широкополосного доступа займет около 11 миллиардов лет
Зеттабайт	Йоттабайт	Весь интернет содержит 1 йоттабайт данных

* Эксабайт — единица измерения количества информации, равная 10^{18} байт. *Прим. ред.*

** Зеттабайт — единица измерения количества информации, равная 10^{21} байт. *Прим. ред.*

Слияние аналитической среды со средой данных

Прежде для проведения анализа специалистам требовалось собрать в одной среде все известные и относящиеся к анализируемому предмету данные. Другого выбора не было. Большая часть аналитической работы, выполняемой профессионалами, относится к сфере передовой аналитики, которая включает в себя интеллектуальный анализ данных, прогнозное моделирование и другие передовые методы. Мы расскажем о них более подробно в главе 7.

Существует любопытная параллель между тем, чем занимались аналитики в самом начале, и тем, что собой представляют хранилища данных. Многих людей удивляет это. Аналитики в течение многих лет работали с «наборами данных». Между «набором данных» и «таблицей» в базе данных существует не так уж много различий. И набор данных, и таблица содержат строки и столбцы. Часто одна строка представляет некую сущность, например потребителя. Столбцы включают в себе информацию о потребителях, например имя, уровень расходов или статус.

Аналитики годами производили «слияние» наборов данных, а это то же самое, что «соединение» таблиц в базе данных. В обоих случаях происходит объединение двух или более наборов данных или таблиц. Как правило, это делается для того, чтобы объединить конкретные строки одного набора данных или таблицы со строками другого набора или таблицы. В нашем примере с потребителем один набор данных / таблица может содержать демографические данные, а другой — информацию о расходах. При объединении наборов данных / таблиц демографические данные и сведения о расходах по каждому потребителю оказываются в одном и том же месте.

Аналитики занимаются тем, что называется «подготовкой данных». Этот процесс подразумевает сведение воедино всех данных из различных источников для создания необходимых для проведения анализа переменных. В случае с хранилищами данных этот процесс называется «извлечение, преобразование и загрузка данных (ETL)». По сути, аналитики создавали пользовательские витрины данных и мини-хранилища данных еще до того, как появились термины «витрины данных» или «хранилища данных»! Просто они это делали при работе над каждым проектом для собственного пользования, а не для использования другими.

Двадцать лет назад большинство аналитиков производили анализ с помощью мейнфреймов. Все данные хранились на больших катушках с пленкой. Я до сих пор помню, как звонил в центр, когда приближались сроки сдачи проекта, с просьбой присвоить моим задачам более высокий приоритет и загрузить мои ленты быстрее. Со временем произошел значительный сдвиг в сторону систем управления реляционными базами данных.

Системы управления реляционными базами данных (СУРБД) начали становиться не только популярными, но и более масштабируемыми и широко применяемыми. Сегодня СУРБД — стандарт управления данными. Большая часть данных, используемых в настоящее время в процессе анализа, берется из реляционных баз данных. Исключение составляет обработка неструктурированных данных с помощью парадигмы MapReduce, о которой говорится далее в этой главе в разделе под названием «Модель MapReduce».

Мощь централизации

Тенденция использования централизованных корпоративных хранилищ данных сильно повлияла на сферу передовой аналитики. Когда данные находятся в одном и том же месте, нет необходимости заниматься их объединением перед проведением анализа. Данные уже ждут, чтобы их проанализировали! Это обеспечивает новый уровень масштабируемости и огромные возможности.

Поначалу базы данных создавались для каждой конкретной цели или команды, а реляционные базы данных распространялись по всей организации. Такие одноцелевые базы данных часто называют «витринами данных». Хотя многие организации до сих пор используют витрины данных, ведущие организации предпочитают объединять различные системы баз данных в одну большую систему, называемую корпоративным хранилищем данных (КХД).

Цель КХД заключается в том, чтобы собрать все важные корпоративные данные в одной центральной базе данных. Это дает возможность объединить различные сведения, что позволяет создавать отчеты и смешивать данные, относящиеся к разным темам или предметным областям. Данные, касающиеся маркетинга, больше не отделены от финансовых данных.

Вот где начинается самое интересное! Как только все данные оказываются в одном и том же месте, больше не существует разнородных источников, которые необходимо объединять для проведения анализа.

Кроме того, там, где собраны все данные, можно проводить более подробный анализ. Рис. 4.1 и 4.2 иллюстрируют разницу между старым и новым способами ведения дел.



Рис. 4.1. Традиционная архитектура аналитической системы

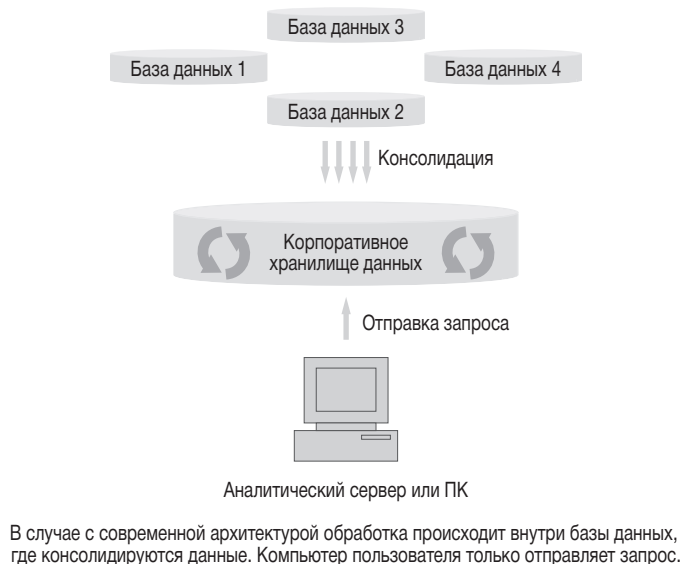


Рис. 4.2. Современная архитектура аналитической системы

В среде корпоративного хранилища данных большая часть источников данных уже собрана в одном месте. Если некоторые мелкие фрагменты данных отсутствуют в КХД, то нет смысла собирать 90–95% данных, находящихся в корпоративном хранилище данных, и сопоставлять их

в режиме офлайн с остальными 5–10%. Гораздо целесообразнее поместить остальные 5–10% данных в КХД и проанализировать все данные там. Другими словами, проще наладить процесс анализа там, где находятся данные, вместо того чтобы перемещать данные туда, где находятся средства анализа. В этом заключается суть концепции *аналитики, встроенной в базу данных*.

Модернизируйте архитектуру

Зачем тратить время, усилия и деньги на перемещение данных в среду для анализа, когда можно переместить средства анализа в среду данных? Эта простая предпосылка, лежащая в основе концепции аналитики, встроенной в базу данных, оказывает огромное влияние на масштабируемость. Без использования аналитики, встроенной в базу данных, укротить большие данные намного сложнее.

Корпорация Teradata первой стала развивать *встроенную в базу данных аналитику* в 1990-х годах, однако сегодня эту концепцию используют все поставщики баз данных. Корпоративные хранилища данных и витрины данных на сегодняшний день достаточно масштабируемы и гибки, чтобы обеспечить возможность проведения анализа в базе данных. Это особенно актуально для массивно-параллельных систем обработки, о которых пойдет речь далее. Ключевая идея, как уже говорилось, состоит в перемещении средств анализа в среду данных вместо перемещения данных в среду для проведения анализа. Позвольте системе баз данных делать то, что она делает лучше всего, — обрабатывать большие массивы информации.

Студенты университетов сегодня мало что знают о мейнфреймах и не могут представить себе способа проведения анализа с использованием ленточных накопителей. А вскоре студентам будет непонятно, почему среда для проведения *углубленного анализа* и среда данных когда-то были разделены. Для них не окажется различий между средами хранения и анализа данных. Эти среды станут представлять для них одно и то же. Так и должно быть!

Массивно-параллельные системы обработки

Массивно-параллельные системы обработки (massively parallel processing, MPP) данных существовали на протяжении десятилетий.

Хотя архитектуры отдельных поставщиков могут варьироваться, массивно-параллельная обработка — наиболее развитой, проверенный и широко используемый механизм хранения и анализа больших объемов данных. Так что же собой представляет массивно-параллельная архитектура и что в ней особенного?

При использовании массивно-параллельной архитектуры данные разделяются на фрагменты, обрабатываемые независимыми центральными процессорами (CPU) и хранящиеся на разных носителях. Это похоже на загрузку разных фрагментов данных на несколько объединенных в сеть персональных компьютеров. Таким образом устраняется ограничение, обусловленное наличием одного центрального сервера с одним процессором и диском. Данные в массивно-параллельной системе распределяются по нескольким дискам, управляемым процессорами разных серверов (рис. 4.3).



Рис. 4.3. Массивно-параллельная система обработки данных

В чем преимущество такой архитектуры? Представьте себе движение по шестиполосному шоссе. Если эти шесть полос сойдутся в одну, пусть даже на коротком участке дороги, движение будет сильно затруднено. Если шесть полос остаются открытыми на всем протяжении пути от отправной точки до места назначения, то поездка будет гораздо более комфортной. В часы пик на дороге могут возникать пробки, но они будут меньше и очень скоро рассосутся. В случае с традиционной архитектурой базы данных в процессе обработки

существует по крайней мере несколько точек, в которых количество полос сокращается до одной. Одной полосы может быть достаточно, только если объем движения небольшой. Именно это делает архитектуру MPP незаменимой для анализа больших объемов данных: она позволяет всем полосам оставаться открытыми на протяжении всего процесса.

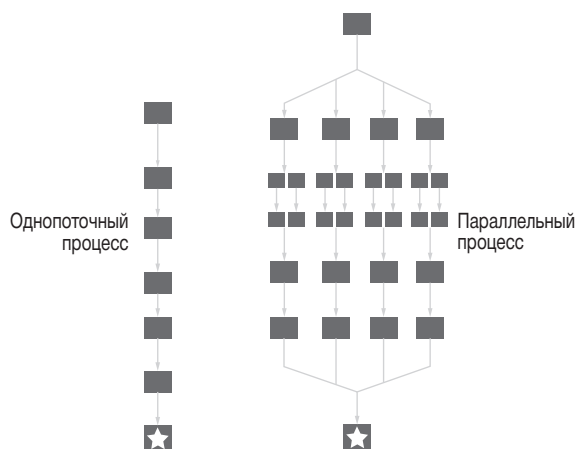
Рассмотрим пример из мира баз данных. Традиционная база данных будет опрашивать терабайтную таблицу по одной строке за раз. Однако при использовании массивно-параллельной системы с 10 обрабатывающими устройствами данные разбиваются на 10 независимых фрагментов по 100 гигабайт. Это означает, что одновременно выполняется 10 запросов. При необходимости в большей вычислительной мощности и более высокой скорости просто добавьте дополнительные обрабатывающие устройства.

Разделяйте работу!

При использовании массивно-параллельной системы данные распределяются по разным процессорам и дискам. Подумайте о десятках или сотнях персональных компьютеров, каждый из которых содержит фрагмент большого набора данных. Это обеспечивает значительно более быстрое выполнение запроса, поскольку вместо одного большого запроса одновременно обрабатывается множество мелких независимых запросов.

Если бы система в нашем примере включала 20 обрабатывающих устройств, то вместо 10 фрагментов по 100 гигабайт одновременно обрабатывалось бы 20 независимых фрагментов по 50 гигабайт, что увеличило бы производительность. Дело несколько усложняется, когда выполнение запроса требует перемещения данных между процессорами, однако массивно-параллельные системы очень быстро справляются и с этой задачей (см. рис. 4.4).

При использовании MPP-систем данные не хранятся в одном месте, что облегчает восстановление в случае выхода оборудования из строя. Эти системы также предусматривают инструменты управления ресурсами для управления процессорами и дисковым пространством, оптимизаторы запросов и другие средства, которые облегчают и повышают эффективность использования систем. Более подробное изложение этой темы выходит за рамки данной книги.



Массивно-параллельная система вместо использования однопоточного процесса обработки данных разделяет работу на части и позволяет различным процессорам и дискам производить обработку одновременно.

Рис. 4.4. Сравнение традиционного и MPP-запроса

Использование MPP-систем для подготовки данных и скоринга

Причина, по которой массивно-параллельная система может оказаться столь полезной в сфере передовой аналитики, заключается в том, что большая часть неудобств, связанных с обработкой, приходится на этап подготовки данных, которая включает соединение, агрегирование, деривацию и преобразование. Соединение подразумевает объединение различных источников данных для сбора всей необходимой для анализа информации. Агрегация предполагает комбинирование сведений из нескольких записей. Один из примеров — вычисление общего и среднего объема продаж, приходящегося на одного клиента, по нескольким транзакциям. К деривации и преобразованию относятся такие действия, как вычисление соотношений, например объем продаж за одну транзакцию, и применение функций, например логарифм или квадратный корень, к переменным для создания дополнительных переменных.

Логика большинства задач, связанных с подготовкой данных, относительно проста. Для решения именно таких задач и предназначены реляционные базы данных и их родной язык, известный как язык структурированных запросов (structured query language — SQL). Совре-

менный SQL может справиться с большинством, если не со всеми, задачами подготовки данных, необходимыми для анализа. Именно использование SQL положило начало развитию *аналитики, встроенной в базу данных*, в MPP-системах. Аналитики просто поместили задачи в базу данных, написав их на SQL, вместо того чтобы писать их на аналитическом языке, что потребовало бы извлечения информации из базы данных.

Еще 10 лет назад SQL имел некоторые ограничения, когда речь шла об определенных сложных вычислениях, необходимых для поддержки передовой аналитики. Сегодня это язык гораздо более мощный. Старое правило, согласно которому при работе с конкретной строкой запрос не учитывает данные в других строках, уже не действует. Например, существуют SQL-функции, называемые «оконными» (windowed aggregates), которые при обработке конкретной строки позволяют запросу учесть данные, находящиеся в другом месте. При помощи таких функций запрос, например, может узнать, является ли данная транзакция первой или последней для клиента, и предпринимать разные действия при обработке данных. Такая функциональность позволяет SQL решать множество дополнительных сложных задач обработки, которые ставят средства передовой аналитики в процессе подготовки данных.

До того как SQL обзавелся более широкой функциональностью, для обработки информации необходимо было извлекать ее из базы данных. К счастью, все это в прошлом, поскольку язык SQL значительно эволюционировал. Большинство операций, связанных с подготовкой данных, теперь могут быть произведены с помощью SQL прямо в базе данных. Многие специалисты до сих пор пишут код на SQL вместо использования аналитических инструментов, однако в настоящее время существует и множество встроенных вариантов, о которых мы расскажем далее в этом разделе.

Не стоит недооценивать SQL

За последние несколько лет язык SQL так эволюционировал, что с его помощью можно решить практически любую задачу, связанную с подготовкой данных. Используя SQL и аналитические инструменты, позволяющие проводить обработку в базе данных, аналитики могут значительно увеличить масштаб процессов. При работе с большими данными требуется больший масштаб, чем когда-либо.

По мере развития концепции обработки информации в базе данных поставщики аналитических инструментов начали дополнять свои

приложения набором функций, позволяющих проводить обработку непосредственно в базе данных. Код написан на родном для этих инструментов языке, однако программное обеспечение теперь распознает осуществление доступа к движку MPP-базы данных, после чего оно передает базе данных инструкции и позволяет ей производить обработку информации вместо ее извлечения.

Развитие аналитических инструментов, обеспечивающих возможность обработки в базе данных, означает, что специалисты теперь могут работать в удобной для них среде и при этом пользоваться преимуществами масштабируемости. По мере увеличения количества функций аналитических приложений, работающих в MPP-базах данных, влияние концепции аналитики, встроенной в базу данных, будет только усиливаться.

Обработка в базе данных применяется при проведении скоринга. Модели часто строятся на основе выборки, однако скоринг подразумевает работу со всеми данными. Например, модель склонности может быть построена на основе выборки клиентов, однако, когда придет время для использования модели, необходимо будет оценить каждого клиента, чтобы выбрать среди них тех, кто наберет наибольшее количество баллов, и нацелить на них маркетинговую кампанию. В таких случаях старые методы извлечения всей информации из базы данных могут быть малопригодными, а то и невозможными из-за временных затрат.

Рассмотрим более подробный пример. Допустим, розничный торговец построил модель склонности, чтобы выявить клиентов, которые, скорее всего, отреагируют на рекламную акцию. Такая модель часто строится на основе репрезентативной выборки, включающей несколько сотен тысяч потребителей. Те, кто отреагировал на специальное предложение в прошлом, сравниваются с теми, кто этого не сделал. Модель создает алгоритм скоринга, который будет вычислять вероятность отклика каждого клиента.

При построении модели извлечение информации из базы данных не связано с особыми сложностями, поскольку представляет собой одноразовое действие и подразумевает работу только с выборкой. Когда придет время для использования модели, алгоритм скоринга будет применен к каждому из десяти миллионов клиентов, чтобы точно определить тех, кто откликнется на предложение с большей или меньшей вероятностью. Этот процесс скоринга будет проводиться регулярно.

Поскольку при этом анализируется вся совокупность потребителей, извлечение данных может свести производительность на нет. Избежать этого можно, если производить обработку в базе данных.

На сегодняшний день существует не менее четырех основных способов проведения подготовки данных и скоринга в базе данных: 1) модель проталкивания SQL; 2) функции, определенные пользователем (UDF); 3) встроенные процессы и 4) язык разметки для прогнозного моделирования (PMML).

Модель проталкивания SQL

SQL — родной язык для MPP-системы, и он отвечает широкому спектру требований. Это особенно касается агрегирования, соединения и преобразования данных. Многие основные задачи, связанные с подготовкой данных, могут быть переведены на язык SQL пользователем. Либо аналитический инструмент может генерировать SQL-код от имени пользователя и «протолкнуть» его в базу данных. Код SQL также легко генерируется многими аналитическими алгоритмами, обладающими довольно простой логикой. К этой категории относятся линейная регрессия, логистическая регрессия и деревья решений. Аналитические инструменты часто переводят логику модели на язык SQL. Иногда пользователи самостоятельно пишут SQL-сценарий после создания модели. В любом случае подготовка данных или процессы скоринга в конечном счете выполняются только с использованием SQL.

Функции, определенные пользователем

Функции, определенные пользователем (UDF), — относительно новая особенность реляционных баз данных. Возможности UDF выходят за рамки возможностей языка SQL. Определенные пользователем функции расширяют функциональность языка SQL, позволяя пользователю определить логику задачи, которая выполняется так же, как и родная SQL-функция.

Запрос общего объема продаж по каждому потребителю может быть записан следующим образом:

«Select Customer, Sum(Sales)...»

При использовании определенных пользователем функций запрос на проведение скоринга потерь может быть записан так:

«Select Customer, Attrition_Score...»

В последнем примере «Attrition_Score» — определенная пользователем функция, имеющаяся в системе базы данных. UDF применяет к данным необходимую логику, и эта логика может быть значительно более сложной, чем в случае использования только языка SQL.

Функции, определенные пользователем, пишутся на таких языках, как C++ или Java. В результате в них могут быть встроены некоторые процедурные возможности языка. Это расширяет ядро SQL, которое не имеет таких возможностей. Сложность заключается в том, что многие профессиональные аналитики не умеют писать программы на этих языках. И тогда на помощь приходят аналитические инструменты, которые автоматически генерируют соответствующие определенные пользователем функции и загружают их в базу данных, подготавливая к использованию.

Встроенные процессы

Концепция встроенных процессов — абсолютно новый способ обработки в базе данных. Встроенный процесс представляет собой еще более высокий уровень интеграции, чем определенные пользователем функции. Последние подразумевают компиляцию кода в виде новых функций, которые могут быть вызваны из запроса SQL, как и любые другие. С точки зрения пользователя аналитическая функция будет вести себя так же, как исходный код аналитического инструмента, и эффективно работать в базе данных, однако она больше не представлена в виде кода аналитического инструмента. По сути, аналитический инструмент перевел то, что ему было необходимо, на язык базы данных.

Тем не менее встроенный процесс представляет собой вариант работы движка аналитического инструмента в самой базе данных. Таким образом, встроенный процесс может запускать программы непосредственно внутри базы данных. Встроенные процессы используют код аналитического инструмента, который был помещен в движок базы данных. Когда необходимо выполнить фрагмент кода и использовать преимущества параллелизма системы, этот фрагмент передается аналитическим инструментам, работающим на каждом из процессоров базы данных. В переводе нет необходимости. Этот метод требует наименьших изменений в исходном коде, однако реализовать его сложно. Поставщики инструментов и баз данных только начинают расширять сферу применения встроенных процессов. В ближайшие годы этот метод, вероятно, станет самым распространенным.

Язык разметки для прогнозного моделирования

Язык разметки для прогнозного моделирования (predictive modeling markup language, PMML) — способ передачи результатов моделирования от одного инструмента другому. Концептуально он вмещает минимальное количество информации, необходимой для создания фрагмента кода в целях скоринга. Информация, включаемая в PMML-поток, содержит тип модели, имена переменных и их форматы, а также значения параметров⁷. Язык PMML позволяет аналитикам использовать любой из PMML-совместимых инструментов для построения модели. Если инструмент PMML-совместим, а для проведения скоринга предполагается использование другого PMML-совместимого инструмента, просто передайте код PMML от первого инструмента второму, и второй инструмент автоматически сгенерирует процесс скоринга.

Один из недостатков PMML проявляется не сразу. В системе, где применяется код PMML, должны присутствовать такие же переменные в том же формате, как и в системе, где была построена модель и сгенерирован PMML-код. Предположим, что в разработанной модели применялась входная переменная «SumOfSales», содержавшая общий объем продаж по каждому клиенту в числовом формате. Переменная «SumOfSales» также должна присутствовать в виде числовой переменной в системе, где будет использоваться код PMML. Для этого в данной системе ее необходимо воссоздать.

Изначально специалисты считали, что благодаря языку PMML им не придется беспокоиться о работе в базе данных в процессе построения моделей, поскольку они могут построить модель, используя аналитический инструмент, а потом с помощью PMML использовать ее в реляционной базе данных. Проблема заключается в требовании наличия одних и тех же метрик в одних и тех же форматах. Если вне базы данных производятся какие-либо манипуляции, то перед применением PMML-кода они должны быть воссозданы внутри базы данных для поддержки процесса скоринга. Код PMML не учитывает какие-либо подготовительные задачи, а только алгоритм скоринга, применяемый к подготовленным данным. Использование кода PMML предполагает, что подготовка данных уже завершена.

Язык PMML еще больше усиливает необходимость в максимизации объема обработки, производимой в базе данных. Для обеспечения максимальной эффективности кода PMML необходимо указать данные, которые уже подготовлены для использования в процессе моделирования

с помощью аналитического инструмента. Данные следует не изменять, а просто обработать с помощью нужных алгоритмов. После этого код для проведения скоринга, сгенерированный PMML-кодом модели, будет готов к немедленному запуску без необходимости в реинжиниринге задач подготовки данных в среде развертывания.

Технологию PMML следует понимать правильно

Язык PMML фактически усиливает необходимость и преимущества проведения аналитической подготовки в базе данных. Если некоторые подготовительные шаги были выполнены с помощью аналитического инструмента перед разработкой модели, то необходимо воссоздать эти манипуляции в базе данных, чтобы обеспечить работоспособность кода PMML. Чем дважды производить подготовку данных в двух средах, лучше сразу провести ее в базе данных.

Более поздние версии языка PMML дают возможность решить некоторые задачи подготовки данных с его помощью, однако на то, чтобы решить все изложенные здесь проблемы, уйдет еще много времени. Из-за его ограничений язык PMML останется наименее часто используемым вариантом.

Использование MPP-систем для подготовки данных и скоринга. Подведение итогов

Платформы массивно-параллельной обработки представляют собой развивающийся и чрезвычайно важный аспект современной аналитической архитектуры. Большинство крупных организаций в настоящее время используют корпоративные хранилища данных, которые содержат огромный объем важных корпоративных данных. Менее крупные компании часто используют реляционные витрины данных. Все больше и больше задач, связанных с обработкой данных, решаются в хранилище данных, и эта тенденция будет развиваться.

Любой организации, которая стремится улучшить масштабируемость аналитической системы, следует обратить внимание на сферу MPP. В связи с увеличением объема данных перемещать их в процессе анализа нецелесообразно, за исключением тех случаев, когда это абсолютно необходимо. Дополнительная масштабируемость также значительно увеличивает широту и объем аналитических процессов,

которые организация может использовать. Эти дополнительные процессы применимы к традиционным данным, большим данным или к их комбинации.

Перед тем как завершить этот раздел, мы должны остановиться на следующем. Хотя мы касались в основном корпоративных хранилищ данных, большинство поставщиков MPP-систем сегодня предлагают «устройства», которые представляют собой уменьшенные версии их систем корпоративного класса. Они предназначены для специальных целей, например для команды аналитиков, которым необходимо обрабатывать огромные объемы данных. Устройства часто бывают созданы для обработки одного или двух рабочих потоков, а корпоративное хранилище данных поддерживает многопоточность.

Углубленная аналитика — один рабочий поток. Если вы планируете использовать корпоративное хранилище данных для углубленной аналитики, удостоверьтесь, что оно поддерживает выполнение этих процессов одновременно с другими процессами, такими как запросы и отчетность. Если нет, рассмотрите возможность использования отдельного устройства. Некоторые производители предлагают устройства, специально предназначенные для профессиональных аналитиков. Они доступны по цене и базируются на тех же принципах, что и корпоративные MPP-системы.

Облачные вычисления

В настоящее время концепции облачных вычислений уделяется много внимания. Как и многие другие технологии, облачные вычисления развиваются в соответствии с циклом зрелости технологий. Для начала определим, что собой представляют облачные вычисления и как они могут быть использованы в сфере передовой аналитики и больших данных. Существует множество определений облачных вычислений. Мы рассмотрим два из них. Первое предложено в докладе компании McKinsey в 2009 году⁸, в котором указаны три критерия для облачной среды:

1. Предприятия не несут капитальных расходов или расходов на инфраструктуру, зато несут эксплуатационные расходы без каких-либо контрактных обязательств.

2. Емкость может быть увеличена или уменьшена динамически и немедленно. Это отличает облака от традиционных хостинг-провайдеров, которые могут ограничивать масштабирование.
3. Базовое оборудование может находиться где угодно. Пользователь абстрагируется от архитектурной спецификации. Кроме того, аппаратные средства работают в многопользовательском режиме, при котором несколько пользователей из разных организаций одновременно могут получить доступ к одной и той же инфраструктуре.

Настоящее облако должно отвечать всем трем перечисленным критериям. Оно должно скрывать базовую инфраструктуру от пользователя, позволять производить масштабирование по требованию и подразумевать оплату по мере использования.

Прыгать? Насколько высоко?

При использовании облака нет необходимости беспокоиться о ресурсных ограничениях. Пользователи могут своевременно получить то, что им необходимо. Они, конечно, заплатят, но только за то, что использовали. Им больше не придется воевать с системными администраторами за ресурсы. Когда облако попросят, чтобы оно прыгнуло, оно спросит: «Насколько высоко?» — вместо того чтобы спорить о том, следует ли ему прыгать вообще!

Другое определение предложено Национальным институтом стандартов и технологий (NIST), входящим в Министерство торговли США. Оно опирается на пять основных характеристик облаков облачной среды⁹:

1. Самообслуживание по требованию.
2. Широкий доступ к сети.
3. Создание пула ресурсов.
4. Быстрая эластичность.
5. Измеряемый сервис.

Каждый из этих критериев должен быть учтен. Легко заметить сходства между определениями McKinsey и NIST. На сайте NIST вы можете больше узнать о работе, проделанной этим институтом в сфере облачных вычислений¹⁰.

У облачных вычислений, как у любого явления, есть хорошие и плохие стороны, плюсы и минусы, преимущества и недостатки. Чтобы сделать осознанный выбор, организациям необходимо получить достаточно информации. Несомненно, в будущем облако будет находить все более широкое применение в сфере углубленной аналитики. Это особенно актуально в сфере разработки; будущее сферы производства менее ясно. Теперь рассмотрим два основных типа облачных сред: 1) публичные облака и 2) частные облака.

Публичные облака

Публичным облакам уделяется наибольшее внимание. При работе с публичным облаком пользователи просто загружают свои данные в хост-систему, а затем, когда им нужно будет использовать эти данные, им будут выделены необходимые ресурсы. И они заплатят за то, что использовали.

Такой подход обладает определенными преимуществами:

- ▶ Обеспечивается необходимая пропускная способность, а пользователи платят только за то, что они использовали.
- ▶ Нет необходимости покупать систему, рассчитанную на максимальную нагрузку, которая большую часть времени будет использовать только половину мощности.
- ▶ Короткие периоды, когда требуется обрабатывать большой объем данных, не вызовут каких-либо хлопот. Нужно просто заплатить за использование дополнительных ресурсов.
- ▶ Нарастить мощности можно, как правило, очень быстро. После получения доступа к облачной среде пользователи загружают свои данные и приступают к анализу.
- ▶ Можно легко обмениваться данными с другими людьми независимо от их местонахождения, поскольку публичное облако по определению находится за пределами корпоративного брандмауэра. Любой человек может получить доступ к созданной среде.

У публичного облака есть и некоторые недостатки:

- ▶ Обычно при использовании публичного облака существует мало гарантий хорошей производительности. Любое количество людей

может одновременно претендовать на использование одних и тех же ресурсов. Конечно, можно доплатить за эксклюзивное использование облачного сервера.

- Это может привести к высокой неустойчивости уровня производительности. Скорость выполнения задачи не будет известна до ее отправки. Можно делать некоторые прогнозы на основании прошлых показателей производительности, однако нет гарантий того, что они будут реализованы.
- Существуют опасения относительно безопасности данных. Хотя есть люди, которые утверждают, что в реальности проблемы менее серьезны, чем в теории, вопросы безопасности представляют собой серьезную проблему.
- Если облако использовать нерационально, это может дорого обойтись, поскольку пользователи платят за все, что они делают. Некорректные запросы, которые используют значительные системные ресурсы, раздражают и в собственной системе, однако с ними не связаны реальные прямые затраты. При использовании облака из-за некорректного запроса вы можете понести довольно большие расходы.
- Работая в публичном облаке, невозможно проверить аудиторский след и местоположение данных. Бывает, что невозможно даже определить, находятся ли все данные в одной и той же стране.

С учетом этих плюсов и минусов каковы целесообразные и нецелесообразные методы использования публичного облака?

Аппаратные ресурсы облака описываются как эластичные, то есть они в любое время могут довольно легко увеличиваться и сокращаться. Это означает легкость наращивания вычислительной мощности за счет использования дополнительных процессоров, систем хранения и объема памяти сервера. Кроме того, в любое время можно легко оплатить и использовать 10 дополнительных серверов. Однако этот вид масштабируемости отличается от массивно-параллельных систем. Большинство серверов публичного облака работают независимо друг от друга. МРР-система представляет собой единую большую систему. Это означает, что, если организация управляет множеством малых и средних процессов, облако может быть для нее большим

подспорьем. Тем не менее облако мало чем сможет помочь, когда речь заходит об управлении процессами, которые выходят за рамки масштабов отдельных серверов, работающих в облачной среде. Хотя программное обеспечение массивно-параллельной системы может работать в облаке, тот факт, что базовое оборудование неизвестно и может измениться в любое время, создает многочисленные проблемы, связанные с производительностью программного обеспечения массивно-параллельной системы.

Вероятно, наилучшее применение публичных облаков — работа, связанная с исследованием и разработкой, при которой не приходится беспокоиться о нестабильности уровня производительности. Если аналитикам необходимо поэкспериментировать с некоторыми новыми данными, чтобы понять, чем они могут быть полезны, то им следует рассмотреть возможность использования облачной среды. В большинстве случаев при проведении аналитических исследований и разработок производительность не представляет проблемы вплоть до развертывания процесса. Для развертывания некритичных аналитических процессов облако — вполне подходящая среда.

Если вы беспокоитесь о защите данных, то публичное облако может создавать проблемы. При использовании публичного облака необходимо применять хорошие протоколы безопасности и инструменты, а также обеспечивать надежную защиту своей среды. Единственное, что находится вне зоны вашего контроля, — это сотрудники компании-поставщика облачного сервиса. Если среди них есть хакеры или воры, они могут нанести такой же ущерб, как и сотрудники внутри компании. Такая вероятность крайне низка, однако нарушение безопасности сотрудником компании-поставщика облачного сервиса вызовет гораздо более сильный общественный резонанс, чем если бы это был сотрудник самой компании. Размещение чувствительной персональной информации в публичном облаке требует принятия мер для обеспечения безопасности. В противном случае не делайте этого.

Вы уверены в том, что экономите деньги?

Если речь идет только об одном аналитике, то его затраты при использовании облака, вероятно, будут гораздо ниже затрат на покупку собственного оборудования. Для больших организаций существует подвох. Когда облаком начинают пользоваться разные люди и отделы, стоимость его использования может превысить стоимость покупки собственной системы!

Существует еще один нюанс, связанный с публичным облаком, о котором люди не задумываются. Действительно, работать с облаком дешевле, чем покупать систему, однако если вам необходимо запускать несколько процессов, то стоимость использования облака может на самом деле превысить стоимость внутрифирменной системы. Чем больше пользователей прибегают к публичному облаку и платят по мере его использования, тем более выгодна покупка внутрифирменной системы. Кроме экономии затрат это может привести и к повышению производительности системы, поскольку организация будет иметь над ней полный контроль.

Со временем публичные облака, вероятно, обеспечат выполнение важнейших функций корпоративного уровня по доступной цене. Поставщики, которые сегодня предлагают более высокий уровень производительности, взимают за дополнительные услуги намного больше, чем за базовые облачные сервисы. Вполне возможно, что поставщики решат проблемы безопасности. До тех пор публичные облака будут использоваться компаниями по большей части для разработки процессов.

Частные облака

Частное облако имеет те же характеристики, что и публичное, однако оно принадлежит одной организации и обычно находится в пределах корпоративного брандмауэра. Частное облако служит тем же целям, что и публичное, но пользоваться им могут только люди или группы в пределах данной организации. На рис. 4.5 приведено сравнение публичного и частного облаков.

Огромное преимущество частного облака состоит в том, что организация полностью контролирует безопасность данных и системы. Данные никогда не выходят за пределы корпоративного брандмауэра, поэтому нет необходимости беспокоиться о том, куда они могут попасть. Данные подвергаются не большему риску, чем в любых других внутрифирменных системах.

Недостаток частных облаков заключается в том, что перед их использованием необходимо приобрести всю инфраструктуру, сводящую в краткосрочной перспективе на нет экономию средств. Капитальные вложения в этом году приведут к увеличению затрат по сравнению с использованием публичного облака. В последующие годы затраты будут ниже, чем при использовании публичного облака. В долгосрочной

перспективе, если облачную среду станут использовать много групп и пользователей, владение частным облаком может быть гораздо более выгодным.

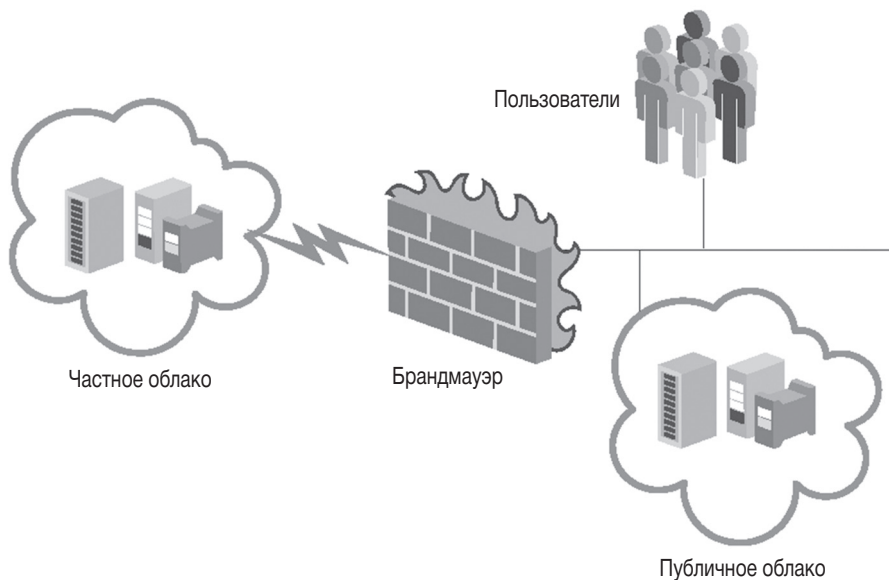


Рис. 4.5. Сравнение публичного и частного облака

Краткосрочные и долгосрочные затраты

В долгосрочной перспективе частное облако может обходиться намного дешевле, чем публичное, если его используют многочисленные пользователи. Тем не менее первоначальные капитальные вложения и настройка потребуют более высоких затрат в краткосрочной перспективе. Со временем баланс будет смещаться.

Концептуально частное облако подобно аналитическим песочницам, о которых идет речь в главе 5. Основное различие заключается в полном самообслуживании частного облака по сравнению с более контролируемой средой песочницы. Тем не менее эти два понятия очень близки, а с практической точки зрения фактически идентичны. Для углубленной аналитики обе среды обеспечивают более или менее одинаковые возможности. При использовании частного облака и строгом соблюдении всех его правил динамическая рабочая нагрузка может привести к снижению производительности, когда пользователи будут бороться за ресурсы. Песочницу можно настроить так,

чтобы по мере необходимости командам выделялись определенные ресурсы. С другой стороны, пользователям песочницы требуется больше усилий, чтобы получить доступ к дополнительным ресурсам, чем пользователям облака.

Облачные вычисления. Подведение итогов

В настоящее время облачным вычислениям уделяется много внимания, и у облачной архитектуры, безусловно, есть преимущества. Организациям необходимо разобраться в способах использования облака, а также понять их плюсы и минусы.

В ближайшее время или в среднесрочной перспективе публичные облака будут применяться в основном для разработки с использованием нечувствительных данных. Частные облака или их близкие «родственники» — аналитические песочницы — станут все чаще применяться для решения всех видов аналитических задач.

Суть в том, что для проведения исследований и содействия инновациям организациям имеет смысл иметь гибкую, слабоструктурированную и менее жестко контролируемую среду. Облака — один из способов создания такой среды.

Грид-вычисления

Некоторые вычисления и алгоритмы невозможно точно преобразовать в код SQL или встроить в определенную пользователем функцию в рамках данных базы. В этих случаях необходимо поместить данные в более традиционную аналитическую среду и применить к ним аналитический инструмент традиционным способом. В течение достаточно длительного времени для решения таких задач использовались крупные серверы. Проблема в том, что по мере увеличения числа специалистов, проводящих анализ, растут количество и размеры серверов. Это связано с большими затратами, а аналитики все равно быстро перегружают доступные вычислительные мощности.

Грид-вычисления могут оптимизировать как стоимость, так и производительность. Данное понятие относится к категории «высокопроизводительных вычислений». Вместо одного высокопроизводительного сервера (или нескольких серверов) используется большое количество менее дорогих компьютеров. В данном случае не сервер распределяет

свои вычислительные ресурсы между заданий, а задания распределяются по разным компьютерам и обрабатываются параллельно. Каждый из этих компьютеров может обладать только частью возможностей исходного сервера и обрабатывать только одно задание за один раз. А в совокупности компьютеры этой грид-сети обрабатывают большие объемы данных. Таким образом, грид-вычисления могут стать экономически эффективным механизмом для увеличения общей пропускной способности и мощности, а кроме того, дадут возможность организации распределять рабочую нагрузку, приоритезировать задания и обеспечивать высокую доступность для аналитической обработки.

Такая сеть позволяет специалистам относительно дешево и быстро масштабировать среду. Однако использование грид-сети не всегда имеет смысл. При обработке нескольких заданий, требующих большой вычислительной мощности, грид-сеть может не являться оптимальным выбором. Поскольку каждое задание приходится на одну машину, обработка больших заданий с помощью дешевых компьютеров может занять гораздо больше времени, чем при использовании более крупного сервера. Тем не менее, если большая организация нуждается в управлении множеством процессов, большинство которых небольшого и среднего размера, грид-сеть может предоставить огромные преимущества.

Последняя инновация в грид-средах — высокопроизводительные аналитические архитектуры, в которых различные компьютеры сети знают друг о друге и могут обмениваться информацией. Это позволяет очень быстро справляться с объемными заданиями, одновременно используя все ресурсы грид-сети, а также решает проблему, связанную с ограничением количества заданий, обрабатываемых одним компьютером сети. Этот вид грид-сетей подает большие надежды и, вероятно, со временем получит широкое распространение. На момент написания этой книги развивается новый вариант, при котором грид-сеть связана непосредственно с системой управления базой данных, что позволяет еще больше увеличить производительность грид-сети. Примером может служить Инфраструктура высокопроизводительной аналитики (High-Performance Analytics) компании SAS.

По мере того как развиваются сложные методы моделирования и такие способы, как товарное моделирование (об этом — в главе 6), продолжают набирать популярность, грид-сеть поможет эффективно справиться с дополнительными рабочими нагрузками, возникающими при моделировании.

Модель MapReduce

MapReduce — это фреймворк для параллельного программирования*. MapReduce не является ни базой данных, ни прямым ее конкурентом. Однако некоторые считают, что эта модель может заменить базы данных и все остальное. Фактически MapReduce представляет собой дополнение к существующим технологиям. Множество задач, решаемых в среде MapReduce, могут быть решены в реляционной базе данных. Все сводится к тому, чтобы выбрать наиболее подходящую среду для решения той или иной проблемы. Если инструмент или технология на что-то способны, то это еще не значит, что этот инструмент или технология самая лучшая. Сосредоточив внимание на том, для каких целей модель MapReduce подходит лучше всего, а не на том, на что теоретически она способна, можно максимизировать полученные преимущества.

Работа MapReduce состоит из двух основных процессов, создаваемых программистом: шага «map» и шага «reduce». Отсюда и название MapReduce. Эти шаги передаются в фреймворк MapReduce, который параллельно запускает программы в наборе рабочих узлов. Вспомните, что при использовании массивно-параллельной системы данные распределяются по узлам, которым затем можно адресовать запрос. В случае с MapReduce используется большое количество недорогого оборудования, на которое по мере необходимости запуска процесса передаются данные. Каждый из рабочих узлов MapReduce применяет один и тот же код к своей части данных. Тем не менее рабочие узлы не взаимодействуют и даже не знают о существовании друг друга.

При наличии постоянного потока данных веб-журналов они могут быть розданы фрагментами различным рабочим узлам. Простейший способ — циклическая процедура (round-robin), при которой записи последовательно снова и снова передаются в узлы. Часто применяются некоторые виды хеширования, в этом случае записи передаются рабочим узлам на основе формулы, чтобы похожие записи отправлялись одному и тому же узлу. Например, при хешировании идентификатора клиента все записи, касающиеся данного клиента, передаются одному и тому же рабочему узлу. Это очень важно, если планируется проведение анализа по идентификатору клиента.

* MapReduce — модель распределенных вычислений, представленная компанией Google, используемая для параллельных вычислений над очень большими (в несколько петабайт) наборами данных в компьютерных кластерах. *Прим. перев.*

Сайт Mapreduce.org определяет модель MapReduce как программный каркас для упрощения обработки массивных наборов данных. Hadoop — популярная версия MapReduce с открытым исходным кодом, поставляемая организацией Apache. Hadoop представляет собой самую известную реализацию фреймворка MapReduce. В данном разделе мы будем использовать общий термин MapReduce, однако вся дальнейшая информация касается любой его реализации.

Организации начинают осознавать важность быстрого анализа огромного количества данных, которые они создают, чтобы принимать более взвешенные решения. MapReduce помогает этим организациям справиться с неструктурированными и частично структурированными источниками, которые сложно анализировать с помощью традиционных инструментов. Большинство предприятий имеет дело с многочисленными типами данных помимо данных из реляционных баз данных. К ним относятся текст, данные, сгенерированные машинами, например веб-журналы, или данные, полученные от датчиков, изображения и т. д. Организациям необходимо быстро и эффективно обрабатывать все эти данные, чтобы извлекать ценные сведения. С помощью модели MapReduce вычислительные операции производятся над данными, хранящимися в файловой системе без загрузки в базу данных. Позднее мы вернемся к этой ключевой особенности.

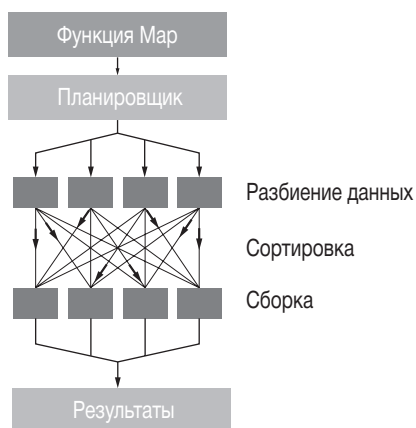
Большое преимущество среды MapReduce заключается в возможности обработки неструктурированного текста. В реляционной базе данных вся информация уже представлена в виде таблиц, состоящих из строк и столбцов. Взаимосвязи между данными уже хорошо определены. Это не всегда верно в случае с потоками необработанных данных. Вот где модель MapReduce действительно может быть полезной! Загрузка больших фрагментов текста в BLOB-поле в базе данных возможна, но это далеко не лучший метод использования базы данных и не лучший способ обработки таких данных. Здесь может помочь модель MapReduce.

Принцип работы MapReduce

Предположим, что есть 20 терабайт данных и 20 серверных узлов MapReduce. Первым шагом будет распределение данных по терабайту на каждый из 20 узлов с помощью простого процесса копирования. Обратите внимание на то, что эти данные должны быть распределены

до запуска процесса MapReduce. Также отметим, что данные находятся в файле определенного пользователем формата. В данном случае нет стандартного формата, как в реляционных базах данных.

Далее программист предоставляет планировщику две программы: «map» и «reduce». При этом двухэтапном процессе программа «map» находит данные на диске и выполняет содержащиеся в ней инструкции. Это происходит на каждом из 20 серверов в нашем примере. Затем результаты выполнения программы «map» передаются процессу «reduce» для агрегации итоговых ответов. На рис. 4.6 этот процесс представлен наглядно.



При использовании модели MapReduce задание разбивается на фрагменты, которые выполняются независимо друг от друга.

Рис. 4.6. Процесс MapReduce

Рассмотрим пример, когда организация получает большой объем текстовых данных из онлайн-чатов отдела обслуживания клиентов на своем сайте. Аналитик создает мар-шаг, чтобы разобрать каждое слово, присутствующее в тексте. В данном примере функция мар будет просто находить каждое слово, отделять его от абзаца и присваивать ему номер 1. В результате получится набор пар значений, например «<мой, 1>», «<продукт, 1>», «<сломался, 1>». После выполнения мар-шага каждый рабочий узел сообщает об этом планировщику.

После окончания мар-шага запускается reduce-шаг. На этом этапе нужно выяснить, сколько раз каждое слово встречается в тексте. То, что происходит далее, называется сортировкой, во время которой

результаты map-шага распределяются с помощью хеширования так, чтобы одни и те же ключевые слова в итоге оказались в одних и тех же узлах. Например, в простом случае существовало бы 26 reduce-узлов, так что все слова, начинающиеся с буквы А, направлялись бы в один узел, слова на букву В — во второй, слова на букву С — в третий и т. д.

Разделяйте работу, часть 2!

Суть MapReduce состоит в том, чтобы разделить бремя обработки большого объема данных среди множества компьютеров. Когда необходимую логику можно применить к различным подмножествам данных, параллельная природа MapReduce позволяет это сделать довольно быстро.

На reduce-шаге происходит просто пересчет слов. В нашем примере результатом проведения reduce-шага будет: «<мой, 10>», «<продукт, 25>», «<сломался, 20>», где числа соответствуют количеству найденных в тексте слов. Будет сгенерировано 26 файлов (по одному для каждого reduce-узла) с отсортированными числами. Следует отметить, что для создания 26 выходных файлов необходим еще один процесс. Чтобы получить итоговый набор ответов, часто требуется несколько процессов MapReduce.

Как только подсчитано количество слов, результаты можно использовать при проведении анализа, например определить частоту появления определенных названий товаров или таких слов, как «сломанный» или «сердитый». Дело в том, что поток текстовых данных, который был совершенно неструктурированным, теперь имеет простую структуру, позволяющую его проанализировать. Использование MapReduce часто служит отправной точкой, а результат применения данного алгоритма используется в качестве входных данных для другого аналитического процесса.

Тысячи процессов MapReduce могут выполняться на тысячах компьютеров. Именно в такой ситуации проявляется вся мощь данной модели. В тех случаях, когда есть огромные потоки данных и процесс их обработки может быть разбит на части, модель MapReduce может применяться с большим успехом. Если для эффективной работы одному рабочему узлу необязательно знать о том, что происходит с другим, можно обеспечить полностью параллельную обработку. В нашем примере каждое слово может быть разобрано само по себе, а содержание других слов не имеет значения для данного рабочего узла.

Предыдущий пункт нельзя игнорировать, поскольку он крайне важен для понимания того, когда и как стоит применять алгоритм MapReduce. Когда данные передаются рабочим узлам, каждый из них знает только о тех данных, которые он видит. Если процесс обработки требует того, чтобы рабочий узел знал о данных, находящихся на других узлах, то следует использовать не MapReduce, а другой фреймворк. К счастью, есть много случаев, когда данные могут быть обработаны таким способом. Разбор одного веб-журнала или одной RFID-записи на фрагменты ни от чего не зависит. Когда необходимо разобрать текст по идентификатору клиента, при распределении данных их просто нужно хешировать так, чтобы все записи, касающиеся данного клиента, оказались на одном и том же рабочем узле.

Концептуально алгоритм MapReduce разбивает проблему на части так же, как это делает параллельная реляционная база данных. Однако MapReduce — это не база данных, ибо у нее отсутствует определенная структура. Каждый отдельный процесс ничего не знает о том, что происходит до или после него. Существуют определенные сходства в том, что вы можете делать с помощью MapReduce и базы данных. База данных может обеспечить входные данные для процесса MapReduce, как и процесс MapReduce — для базы данных. Ключевой фактор — выбор наилучшего способа для решения конкретной задачи. Если что-то можно сделать с помощью данного набора инструментов, это не означает, что этот способ наилучший. Другие наборы инструментов могут быть гораздо более подходящими. Базы данных и модель MapReduce должны использоваться для решения тех задач, для которых они лучше всего подходят.

Сильные и слабые стороны модели MapReduce

Процесс MapReduce работает на стандартном недорогом оборудовании. Это означает, что его можно наладить с небольшими затратами. Кроме того, его расширение обойдется довольно дешево. Увеличить емкость легко, поскольку все, что требуется, — это дополнительные серверы и их интеграция.

Мы уже говорили о том, что с решением некоторых задач MapReduce справляется гораздо лучше, чем реляционные базы данных. Это разбор текста, обработка данных веб-журналов и чтение информации из огромного источника необработанных данных. Модель MapReduce наиболее

эффективна в случаях, когда существует большой объем входных данных, большая часть которых не нужна для анализа. Если большое значение имеет малая часть данных, но заранее неясно, какая именно, то MapReduce может помочь. Этот алгоритм позволяет разобраться в большом объеме данных и извлечь важные фрагменты.

Словно пить из шланга

Многие большие потоки данных, например веб-журналы, содержат большое количество информации, не имеющей долгосрочной ценности. Модель MapReduce позволяет извлекать из потока данных то, что вам нужно, пропуская остальную информацию, как если бы вы пили воду из шланга.

Нет смысла тратить много времени и ресурсов на загрузку огромного количества необработанной информации в корпоративное хранилище данных, если в процессе их обработки большая часть будет отброшена. Если данные нужны только на короткий период, не стоит помещать их в хранилище данных. Модель MapReduce идеально подходит в таких случаях. Отбросьте излишки информации до ее загрузки в базу данных.

Во многих ситуациях модель MapReduce используется как инструмент извлечения, преобразования и загрузки данных (ETL). ETL-инструменты прочитывают набор исходных данных, производят комплекс действий, связанных с форматированием или реорганизацией, а затем загружают результаты в итоговый источник данных. Для обеспечения анализа ETL-инструменты берут данные из систем взаимодействия подразделений предприятия и загружают их в реляционную базу данных, чтобы обеспечить к ним доступ. Модель MapReduce часто используется для обработки источников больших данных, извлечения из них имеющих ценность сведений и передачи результатов в базу данных или аналитический процесс. В предыдущем примере необработанный текст превращается в набор слов с указанием частоты их использования, который можно проанализировать, а результаты передать в базу данных, чтобы объединить эту информацию с дополнительными данными.

Модель MapReduce не база данных, поэтому не имеет встроенной системы безопасности, средств индексирования, оптимизатора запросов или процессов, истории выполненных заданий и сведений о других существующих данных. Она обеспечивает не только максимальную

гибкость при обработке различных видов данных, но также несет ответственность за определение того, что собой представляют данные, в каждом созданном процессе. Практически всегда будет применяться пользовательский код, в том числе и для структуры данных. Каждое задание представляет собой отдельный элемент, которому ничего не известно о том, что происходит за пределами модели.

MapReduce по-прежнему находится на стадии развития. Немногие знают о том, как ее правильно использовать, конфигурировать или писать нужный код. Сегодня уровень компетентности в применении MapReduce сдерживается ресурсными ограничениями. Это будет меняться с течением времени по мере становления данной технологии и роста числа людей, эффективно ее использующих. Тем не менее на момент написания данной книги этой теме уделяется большое внимание.

Модель MapReduce. Подведение итогов

По мере того как организации в своей деятельности все в большей мере смогут опираться на большие данные, модель MapReduce будет завоевывать все большую популярность и влияние. Возможность параллельного запуска процесса на стандартном недорогом оборудовании достаточно заманчива при работе с огромными объемами данных, большая часть которых в долгосрочной перспективе не будет представлять интереса или ценности. Разделение задачи на мелкие фрагменты позволяет решить ее быстрее и дешевле.

MapReduce не база данных и не заменяет ее. Однако эта модель способна существенно увеличить качество баз данных предприятия. После того как MapReduce обработает и извлечет важные фрагменты из потока больших данных, их можно поместить в традиционную среду базы данных для дальнейшего более глубокого анализа, а также для обеспечения более широкого доступа к запросам и отчетам. В некотором смысле модель MapReduce представляет собой более мощную версию ETL-процесса.

Завершим раздел небольшим примером. Веб-журналы содержат огромный объем бесполезных данных. Модель MapReduce может найти ценные иглы в стогу сена. Представьте, что с помощью алгоритма MapReduce журналы обрабатываются практически в режиме реального времени, чтобы определить необходимые меры, например найти всех клиентов, которые просмотрели данные о товаре, но не купили его.

Процесс MapReduce формирует список клиентов, которым необходимо отправить электронное письмо, и эта информация немедленно отправляется процессу, генерирующему такие письма, причем без первоначальной загрузки необработанных данных в реляционную базу данных и выполнения запроса.

После выполнения первоначальной задачи наиболее важные фрагменты данных загружаются в базу данных и пополняют остальную важную информацию о клиенте, что позволяет производить более полный стратегический анализ по периодам и подразделениям организации. В данном примере список выявленных клиентов загружается в базу данных — так учитывается, что им были отправлены электронные письма. Это позволит отследить историю переписки, как это делается при проведении любой e-mail-кампании.

Речь не о выборе или/или!

Массивно-параллельные реляционные базы данных, облака и модель MapReduce играют собственную роль в аналитической экосистеме, которая позволит укротить большие данные. Все три технологии могут быть использованы вместе для достижения максимальных результатов. Существует множество способов их объединения:

- ▶ Разработаны базы данных, работающие в облаке.
- ▶ Существуют базы данных со встроенной функциональностью MapReduce. Например, платформа Aster компании Teradata имеет запатентованную реализацию MapReduce на языке SQL, которая позволяет выполнять процессы MapReduce в качестве части SQL-запроса.
- ▶ MapReduce может применяться к данным, получаемым из базы данных, или быть источником информации для базы данных.
- ▶ MapReduce может применяться к данным в облаке.
- ▶ MapReduce может применяться к данным в базе данных, размещенной на облаке.

Эти три технологии могут взаимодействовать друг с другом и работать совместно. При правильном использовании они дополняют друг друга. Вам не нужно выбирать только одну из них. Они могут быть

частью аналитической экосистемы, и многие организации будут использовать все три технологии. Грид-вычисления также могут применяться по мере необходимости в любом из приведенных сценариев.

Обзор главы

Самые важные уроки этой главы.

- Аналитики на протяжении десятилетий раздвигают границы масштабируемости. Большие данные — всего лишь следующее поколение данных, которые необходимо укротить.
- Аналитическая среда соединяется со средой управления данными. Аналитика, встроенная в базу данных, заменяет большую часть традиционных способов аналитической обработки, используемых для обеспечения углубленной аналитики.
- Массивно-параллельная обработка (MPP), облачные архитектуры и модель MapReduce — мощные инструменты, помогающие в работе с большими данными.
- Профессиональные аналитики могут использовать MPP-базы данных для подготовки данных и скоринга с помощью языка SQL, определенных пользователем функций (UDF), встроенных процессов и языка разметки для прогнозного моделирования (PMML).
- Облака могут быть публичными или частными, они позволяют вам легко получить необходимые ресурсы. Вы платите только за то, что используете. Облака лучше всего подходят для решения задач, связанных с исследованиями и разработкой.
- При использовании публичных облаков уровень производительности не гарантирован; большое внимание следует уделять безопасности, а данные находятся вне зоны непосредственного контроля организации.
- Стоимость работы с публичным облаком может превышать затраты на содержание внутрифирменной инфраструктуры, если оно используется большим количеством сотрудников.
- Частные облака — отличный выбор для крупных организаций, обеспечивающий гибкость в безопасной среде.

- ▶ Грид-сети помогают масштабировать процессы, которые не могут быть переданы в базу данных. Функциональность грид-сети расширяется и становится более мощной.
- ▶ Фреймворк MapReduce — развивающаяся технология, которая позволяет программам работать параллельно.
- ▶ MapReduce помогает укротить большие данные путем проведения предварительной обработки и передачи важных фрагментов информации для дальнейшего анализа.
- ▶ Реляционные базы данных, облака и модель MapReduce играют свою роль в процессе укрощения больших данных. Эти три технологии могут работать совместно и дополнять друг друга для обеспечения большей эффективности.

Эволюция аналитических процессов

Высокая масштабируемость, о которой шла речь в главе 4, не дает организации особых преимуществ, если не используется должным образом. Модернизация технологий не принесет особой пользы, если аналитические процессы останутся прежними. Это напоминает покупку нового 3D-телевизора со всеми «примочками» и его подключение к антенне, ловящей местные ТВ-сигналы. Изображение может улучшиться по сравнению со старым телевизором, однако ваши впечатления от просмотра не сильно изменятся, если учесть все возможности нового оборудования.

Точно так же по мере роста компетентности специалистов, применяющих средства передовой аналитики, должны развиваться аналитические процессы. Унаследованные процессы развертывания аналитических процедур не соответствуют имеющимся возможностям. Без изменения аналитических процессов организации смогут реализовать лишь часть мощности и производительности, которые сегодня доступны благодаря новым уровням масштабируемости. Невозможно укротить большие данные, используя только традиционные подходы к разработке аналитических процессов.

Один из процессов, подлежащих пересмотру, — конфигурирование и поддержание рабочего пространства для аналитиков. Традиционно это рабочее пространство располагалось на отдельном сервере, предназначенном для аналитической обработки. Как уже говорилось, новым стандартом становится аналитика, встроенная в базу данных. Однако для того чтобы воспользоваться преимуществом данного подхода,

необходимо, чтобы рабочее пространство аналитиков, или «песочница», находилось непосредственно в системе управления базой данных. В случае с большими данными среда MapReduce часто является дополнением к традиционной «песочнице». В первой части этой главы мы расскажем, что собой представляет аналитическая песочница, каково ее значение и как ее можно использовать.

По мере того как специалисты с помощью песочницы в процессе своей работы будут использовать платформу базы данных, им придется неоднократно решать определенные задачи. Например, каждому аналитику понадобятся некоторые основные метрики, касающиеся клиентов, независимо от конкретного вида производимого анализа. Ключевым инструментом, помогающим внести в передовые аналитические процессы организации последовательность, обеспечить производительность и снизить риски, являются аналитические наборы данных предприятия. Вторая часть главы начнется с обзора того, что собой представляет простейший аналитический набор данных. Затем мы поговорим о том, что представляет собой аналитический набор данных предприятия (enterprise analytic data set, EADS), каковы его преимущества и как он может применяться другими пользователями или приложениями, не считая специалистов, для которых он разрабатывался.

Многие виды анализа подразумевают «процедуры скоринга», которые следует развернуть и производить на регулярной основе. Например, процедура скоринга модели склонности клиента может оценить вероятность совершения покупки тем или иным клиентом в следующем месяце. Раньше обновление этого показателя для каждого клиента требовало значительного количества времени. В современном мире часто бывает необходимо, чтобы показатели обновлялись ежедневно, а то и в режиме реального времени. В третьей части главы мы расскажем о том, как встроить процедуры скоринга в среду базы данных, а также о способах эффективного отслеживания и управления моделями и процессами, разрабатываемыми с помощью управления моделями.

Аналитическая песочница

В главе 4 мы говорили о преимуществах массивно-параллельных систем. Одно из применений такой системы — содействие созданию и внедрению процессов углубленной аналитики. Однако для более

эффективного использования корпоративного хранилища данных или витрины данных специалистам необходимы особые права и доступ. Аналитическая песочница — механизм, необходимый для их работы. При правильном использовании аналитическая песочница может быть одним из главных факторов, определяющих повышение ценности, в мире больших данных.

Термин «песочница» восходит к песочнице, в которой играют дети. В песочнице они могут создавать все, что им захочется, например слепить что-то из песка. Точно так же песочница в контексте аналитики представляет собой набор ресурсов, которые позволяют специалистам экспериментировать и изменять данные любым способом. Для описания концепции песочницы используются также термины «гибкое аналитическое облако» и «лаборатория данных». Не имеет значения, какой термин вы выберете. Важно, что вы будете использовать данную концепцию.

Аналитическая песочница: определение и сфера применения

Аналитическая песочница предоставляет набор ресурсов, с помощью которого можно произвести глубокий анализ, чтобы ответить на важные бизнес-вопросы. Аналитическая песочница идеально подходит для исследования данных, разработки аналитических процессов, доказательства концепций и прототипирования*. Как только дело доходит до управляемых пользователем или производственных процессов, в применении песочницы отпадает необходимость.

Песочница будет применяться довольно небольшой группой пользователей. В песочнице будут создаваться данные, отделенные от производственной базы данных. Пользователи песочницы также будут иметь возможность загружать собственные данные как часть проекта на короткие периоды, даже если эти данные не включены в официальную модель данных предприятия.

Данные в песочнице будут иметь ограниченный срок актуальности. Идея заключается не в создании массива постоянных данных: во время

* Прототипирование (англ. *prototyping* — первообраз) — реализация базовой функциональности для анализа работы системы в целом, поиска «узких мест». Используется в машино- и приборостроении, программировании, во многих других областях техники. После этапа прототипирования обязательно следуют этапы пересмотра архитектуры системы, разработки, реализации и тестирования конечного продукта. *Прим. ред.*

работы над проектом следует создавать необходимые для него данные. После завершения проекта данные следует удалить. При правильном использовании песочница может стать одним из основных факторов, обеспечивающих дополнительную аналитическую ценность для организации.

Преимущества аналитической песочницы

Каковы преимущества аналитической песочницы? Рассмотрим этот вопрос с точки зрения аналитика и ИТ-специалиста.

Преимущества с точки зрения профессионального аналитика:

- **Независимость.** Профессиональные аналитики смогут работать в системе баз данных без необходимости постоянно запрашивать разрешение для выполнения конкретных проектов.
- **Гибкость.** Профессиональные аналитики смогут использовать любые необходимые аналитические инструменты, будь то средства бизнес-аналитики, статистического анализа или визуализации.
- **Эффективность.** Профессиональные аналитики смогут использовать существующие корпоративные хранилища данных или витрины данных без необходимости в перемещении или переносе данных.
- **Свобода.** Профессиональные аналитики смогут тратить меньше времени на администрирование систем и наблюдение за производственными процессами, передав эти функции обслуживания ИТ-специалистам.
- **Скорость.** Переход к параллельной обработке обеспечит значительное увеличение скорости. Кроме того, это даст возможность совершать большее количество попыток и смелее подходить к внедрению инноваций.

Песочница дает преимущества всем!

Среда песочницы обладает определенными преимуществами как для профессиональных аналитиков, так и для ИТ-специалистов. В данном случае одна группа не выигрывает за счет другой. Люди часто боятся новой концепции, поскольку не понимают ее. Потребуется некоторое время на то, чтобы обучить людей и преодолеть первоначальную реакцию. Однако эти усилия оправдают себя.

Преимущества с точки зрения ИТ-специалиста:

- ▶ **Централизация.** ИТ-специалист сможет централизованно управлять средой песочницы так же, как любой другой средой базы данных в системе.
- ▶ **Оптимизация.** Песочница значительно упростит переход от аналитических процессов к производству благодаря наличию единой платформы для разработки и внедрения.
- ▶ **Простота.** На стадии разработки больше не нужно создавать процессы, которые придется полностью переписывать для запуска в производственной среде.
- ▶ **Контроль.** ИТ-специалист будет в состоянии контролировать среду песочницы, балансировать ее потребности с потребностями других пользователей. Производственная среда не пострадает, если эксперимент в песочнице не удастся.
- ▶ **Затраты.** Путем консолидации многочисленных аналитических витрин данных в единую центральную систему можно обеспечить значительную экономию средств.

Внутренняя песочница

Внутренняя песочница представляет собой выделенную часть корпоративного хранилища данных или витрины данных. Такая песочница физически расположена в производственной системе. Однако сама база данных песочницы не интегрирована в производственную базу данных. Песочница представляет собой отдельный контейнер базы данных в рамках системы (рис. 5.1).

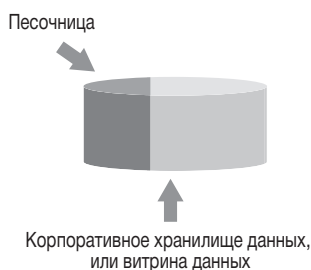


Рис. 5.1. Внутренняя песочница

Обратите внимание: в случае с большими данными целесообразно добавить среду MapReduce. Обычно она устанавливается в дополнение к платформе базы данных, если только вы не используете систему, которая позволяет объединить обе среды. Среда MapReduce потребует обеспечения доступа к внутренней песочнице. Две среды могут обмениваться данными по мере необходимости. О модели MapReduce шла речь в главе 4.

Одна из сильных сторон внутренней песочницы в том, что она может использовать существующие аппаратные ресурсы и инфраструктуру. Это позволяет очень легко наладить ее применение. С точки зрения администрирования нет никакой разницы между созданием песочницы и созданием любого другого контейнера базы данных в системе. Отличия песочницы заключаются в некоторых разрешениях, которые предоставляются ее пользователям, и в способах ее использования.

Вероятно, самое большое преимущество внутренней песочницы заключается в возможности напрямую объединять производственные данные с данными песочницы. Поскольку все производственные данные и все данные песочницы находятся в производственной системе, легко связать эти источники друг с другом и работать со всеми данными одновременно (рис. 5.2).

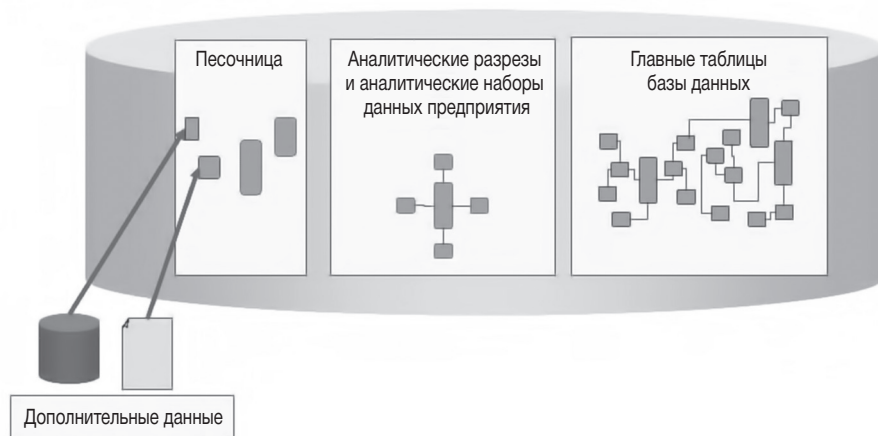


Рис. 5.2. Устройство внутренней песочницы

Внутренняя песочница — очень экономичная технология, поскольку для ее использования не требуется новое оборудование. Производственная система уже существует — она просто используется по-новому. Лик-

видация необходимости в перемещении данных между платформами также снижает затраты. Единственное исключение — необходимость в перемещении данных между базой данных и средой MapReduce.

У внутренней песочницы есть несколько слабых сторон. Одна из них заключается в дополнительной нагрузке на существующее корпоративное хранилище данных, или витрину данных. Песочница будет использовать как дисковое пространство, так и ресурсы процессора (в потенциале — довольно интенсивно). Внутренняя песочница может быть ограничена производственной политикой и процедурами. Например, если в понедельник утром практически все системные ресурсы необходимы для создания отчетов, то пользователям песочницы окажутся доступными лишь минимальные ресурсы.

Внешняя песочница

В случае с внешней песочницей создается физически отдельная аналитическая песочница для тестирования и разработки аналитических процессов. Чисто внешняя среда используется относительно редко. Чаще встречаются внутренние, или гибридные песочницы, о которых мы поговорим далее. Однако важно понимать, что собой представляет внешняя песочница, поскольку она является компонентом среды гибридной песочницы (рис. 5.3).

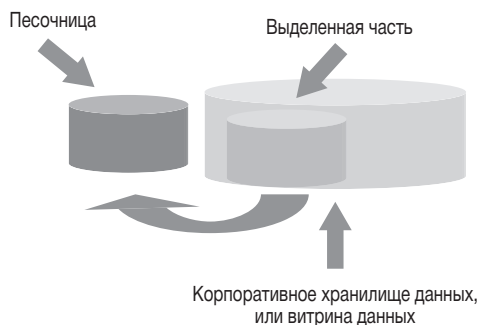


Рис. 5.3. Внешняя песочница

Самое большое преимущество внешней песочницы заключается в ее простоте. Песочница — это автономная среда, предназначенная для разработки процессов углубленной аналитики. Она не влияет на другие процессы, чем обеспечивает гибкость при разработке и использовании.

Например, можно экспериментировать с различными параметрами базы данных или перейти на более новую версию базы данных, чтобы протестировать новый функционал. Так часто делается в традиционных системах тестирования и разработки, используемых для создания приложений.

Часто возникает вопрос: не нарушает ли внешняя система концепцию, согласно которой при проведении анализа данные должны находиться в базе данных? Ответ на этот вопрос — нет, если вы будете рассматривать ее в качестве среды аналитической разработки. Большинство организаций имеют независимую от производственной системы среду тестирования и/или разработки для целей бизнес-аналитики. Это необходимо в процессе создания, тестирования и отладки новых процессов. Внешняя песочница представляет собой точно такую же концепцию по тем же причинам, только предназначена она для аналитических инициатив.

Другое преимущество внешней песочницы состоит в том, что уменьшается необходимость в управлении рабочей нагрузкой. Когда систему используют только аналитики, нет необходимости беспокоиться о настройке и балансировке. Работа среды песочницы и производственной среды будет предсказуемой и стабильной. Например, в понедельник утром пользователи песочницы не столкнутся с дефицитом ресурсов, связанным с созданием отчетов. Они будут обеспечены устойчивым доступом к песочнице.

Внешняя песочница не нарушает правил

Внешняя песочница не нарушает правил обработки в базе данных. Внешнюю песочницу следует рассматривать как среду для тестирования и разработки аналитических процессов. Существует множество веских и убедительных причин для использования таких сред, и они повсеместно применяются для разработки приложений и отчетов.

Внешняя песочница обычно представляет собой реляционную базу данных той же природы, что и производственная система. Таким образом, перемещение процессов из песочницы в производственную среду подразумевает простое копирование. Если извлеченные данные, отправленные в песочницу, хранятся в той же структуре, что и производственные данные, осуществить перенос легко.

Когда дело доходит до работы с большими данными, следует включить MapReduce как часть среды внешней песочницы. В этом случае среда

внешней песочницы будет содержать реляционную базу данных и компонент MapReduce. В одних случаях одна система справится с обеими функциями, в других потребуется две физические платформы.

Главная слабость внешних песочниц заключается в необходимости дополнительных расходов на автономную систему, которая служит платформой для песочницы. В целях экономии многие организации, обновляя свои производственные системы для создания среды песочницы, используют старое оборудование, которое в противном случае было бы выброшено, и это позволяет сэкономить на покупке оборудования для песочницы.

Еще одно слабое место — необходимость перемещения некоторых данных. До разработки нового аналитического процесса в песочницу должны быть перемещены данные из производственной системы; нужно также поддерживать потоки данных. Они могут не быть слишком сложными, однако это дополнительный набор заданий, требующих выполнения. Любые потоки данных следует строго ограничить и фокусироваться только на том, что абсолютно необходимо.

Гибридная песочница

Среда гибридной песочницы — это сочетание внутренней и внешней песочниц. Она позволяет аналитикам при необходимости использовать мощь производственной системы и гибкость внешней системы в целях глубокого анализа или решения задач, не являющихся дружественными для базы данных (рис. 5.4).

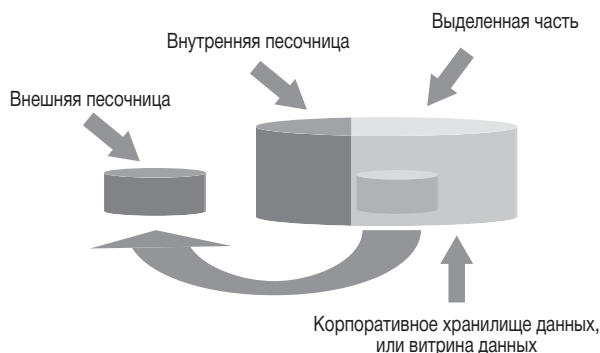


Рис. 5.4. Гибридная песочница

Сильные стороны гибридной среды объединяют преимущества внутренних и внешних песочниц, а также гибкость в выборе подхода

к анализу. При работе во внешней песочнице легко избежать воздействия на производственные процессы на стадии раннего тестирования. Когда приходит время для итогового тестирования и предварительного развертывания, можно использовать производственную песочницу. Среда MapReduce может дополнить гибридную песочницу путем поддержки как внутренней, так и внешней песочницы.

Другое преимущество проявляется, когда созданный аналитический процесс необходимо временно запустить в «псевдопроизводственном» режиме во время полного развертывания производственной системы. Такие процессы легко запустить из внутренней песочницы.

Слабые стороны гибридной среды включают недостатки обоих вариантов, но с некоторыми дополнениями. Одна из слабых сторон заключается в необходимости поддержания среды и внутренней, и внешней песочницы. В этом случае нужно поддерживать согласованность работы не только внешней песочницы и производственной среды, но и внешней песочницы и внутренней песочницы.

В каких случаях следует использовать тот или иной вариант песочницы? Необходимо определить конкретные виды задач, предназначенных для внешней и внутренней песочниц. Специалисты не могут произвольно использовать ту или иную среду. Команда аналитиков должна разработать рекомендации и придерживаться их.

Не переполняйте песочницу

В среду внешней песочницы следует копировать минимальный объем данных, необходимых для анализа. В песочнице должна размещаться только небольшая часть данных, хранящихся в производственной среде. Со временем конкретные данные будут меняться в зависимости от текущих аналитических потребностей. Создавать копии следует только в случае крайней необходимости.

Последний недостаток заключается в том, что могут потребоваться несколько двусторонних потоков данных; это усложнит работу. Данные, доступные для внутренней и внешней песочниц, должны быть согласованными. По мере разработки новых данных в одной из сред может потребоваться воспроизвести их в другой.

Данные нужно не только использовать, но и развивать!

Один из лучших способов применения песочницы — постоянное выявление новых источников данных, которые следует добавить в органи-

зационные системы и процессы. Возможно, вы покупаете поток данных из социальных медиа или файл с демографическими данными или же получаете поток информации из нового источника больших данных. Как аналитики будут изучать эти новые данные и экспериментировать с ними?

Представьте себе, как нерационален был бы типичный подход использования новых данных до их исследования! Вам пришлось бы оправдать и описать проект для загрузки данных. После этого потребовалось бы разработать процессы извлечения, преобразования и загрузки данных (ETL), чтобы загрузить данные в систему. Необходимо было бы разработать, утвердить и реализовать модель данных. Затем все перечисленное пришлось бы протестировать. Через три-шесть месяцев процесс был бы запущен, а данные подготовлены к использованию. В этот момент анализ может показать, что эти данные не имеют большой ценности и вам не нужны. Сколько ресурсов было бы потрачено впустую на формальное добавление этих данных в систему!

Сначала опробуйте образец

Если люди не уверены, понравится ли им вкус мороженого, они его пробуют. Если вкус нравится, они заказывают целую порцию. Если нет, переходят к другому сорту. Следуйте той же логике при работе с новыми источниками данных, особенно с источниками больших данных. Не покупайте целую порцию, если вы не уверены в том, что это именно то, что вам нужно. Сначала поэкспериментируйте с данными в своей песочнице.

Чтобы избежать такого сценария, следует взять фрагмент новых данных, загрузить их в песочницу и протестировать. Если ожидания не оправдались, двигайтесь дальше! Если оправдались, можно начинать длительный и дорогостоящий процесс формального использования данных. Применить аналитическую песочницу для исследования и доказательства ценности новых источников данных значительно быстрее и дешевле, чем использовать традиционные способы.

Управление рабочей нагрузкой и планирование мощностей

Существует множество компонентов систем баз данных, которые обеспечивают надежную работу песочницы. Пользователей песочницы можно отнести к группе, которая имеет разрешение на разработку новых передовых аналитических процессов. Можно, например,

ограничить ресурсы процессора, выделяемые для данного пользователя песочницы. Системы корпоративного класса достаточно гибки, чтобы выделить для пользователей только 10% ресурсов в период высокой нагрузки, однако в ночное время всю систему может использовать один пользователь.

Контролируются количество одновременных запросов или даже типы запросов пользователей. Например, им может быть позволено выполнять лишь пять параллельных заданий одновременно. Могут существовать процессы для выявления и отмены плохо сформированных запросов, например запроса, содержащего перекрестное соединение двух больших таблиц.

Важно ограничить использование дискового пространства посредством политики сохранения данных. Если набор данных в песочнице не был востребован в течение пары месяцев, его следует удалять по умолчанию. Песочницы не должны использоваться для непрерывного наращивания наборов данных, как это часто бывает в традиционных средах.

У некоторых моих клиентов было пять терабайт различных корпоративных данных, однако их аналитическая среда содержала от 30 до 50 терабайт. Причина в том, что каждый аналитик сделал копию большей части данных. У каждого аналитика, возможно, было даже несколько копий данных для разных проектов. Вот почему существует огромное количество избыточных данных. Один и тот же подход не должен повторно использоваться в среде песочницы. Данные в песочнице должны удаляться, если только не существует конкретной причины для их сохранения.

По мере того как во внутренней песочнице запускается все больше аналитических процессов, будут меняться соотношения и уровни использования ресурсов как в среде песочницы, так и в производственной среде. Это нормально. Поскольку среды работают на одной стандартизированной платформе, аналитическая обработка может быть учтена в прогнозах использования ресурсов точно так же, как все остальное. Планы загрузки мощностей следует обсудить перед началом работы, однако в обработке данных в песочнице нет ничего особенного, что могло бы изменить действия людей, которые разрабатывают эти планы. Работа в песочнице просто встраивается в общий процесс. Системные администраторы знают, как это обеспечить.

Бытует распространенное заблуждение, что аналитическая песочница способна «уничтожить» систему, использовать все ресурсы и спровоцировать хаос. Это неправда. Очень крупные аналитические задания,

как правило, необходимо запустить один или два раза в начале работы над проектом. Их не нужно запускать снова и снова. Запуск крупных заданий можно легко запланировать, например, на ночное время, когда система не загружена на полную мощность. Вместо того чтобы затратить все системные ресурсы, аналитическая песочница способна привести к противоположному результату. Запущенные в песочнице аналитические процессы могут использовать ресурсы, которые в противном случае были бы невостребованными. Это позволяет получить дополнительную отдачу от инвестиций в инфраструктуру без лишних расходов. Это здорово!

Истина прямо противоположна тому, во что многие верят!

Среда песочницы может обеспечить создание дополнительной ценности за счет текущих инвестиций, а не за счет дополнительных затрат. Ее использование не подразумевает необходимости в покупке нового оборудования. Кроме того, песочница не мешает другим процессам. Она повышает отдачу от инвестиций без каких-либо негативных последствий. Как только вы поймете, что собой представляет песочница и как она работает, вы осознаете, что истина прямо противоположна тому, во что многие верят!

И последний важный момент. Добавление аналитических процессов в среду песочницы само по себе не требует новых мощностей. Если в настоящее время система используется на 95–99%, то добавление внутренней песочницы, вероятно, потребует обновления системы. Это вызвано только тем, что система настолько загружена, что любое новое приложение или процесс, добавленные в нее, потребуют наращивания мощностей. Точно так же, если для создания внешней песочницы используется старое оборудование, не возникает необходимости в новых затратах. На самом деле дополнительная ценность будет обеспечиваться благодаря оборудованию, которое в противном случае было бы выброшено и не приносило бы никакой пользы.

Что такое аналитический набор данных?

Аналитический набор данных (analytic data set, ADS) — это данные, собранные с целью создания аналитического процесса или модели и представленные в формате, необходимом для решения конкретной аналитической задачи. ADS создается путем преобразования, агрегирования и объединения данных. Он имитирует денормализованную структуру,

или плоский файл. Это означает, что в нем будет присутствовать одна запись для одного клиента, местоположения, товара или любой другой сущности, которую необходимо проанализировать. Аналитический набор данных помогает преодолеть разрыв между эффективным хранением и простотой использования.

Большая часть данных в реляционных базах данных хранится в так называемой третьей нормальной форме. Этот метод хранения данных устраняет их избыточность, но усложняет запросы. Таблицы, которые находятся в третьей нормальной форме, очень эффективны для хранения и извлечения данных, но их нельзя непосредственно использовать в большинстве передовых аналитических процессов. Подробное описание третьей нормальной формы выходит за рамки этой книги. Важно то, что аналитические инструменты, как правило, используют данные в простом, денормализованном виде — в формате плоского файла. Сложность углубленной аналитики заключается в алгоритмах и методах, применяемых к данным, а не в структуре самих данных. Аналитические наборы данных могут принимать различные формы, об этом речь пойдет далее.

Сравнение аналитических наборов данных для разработки и производственных аналитических наборов данных

Существует два основных вида аналитических наборов данных (рис. 5.5). ADS для разработки — это набор данных для создания аналитического процесса. Он включает все переменные, которые могут потребоваться для решения задачи. Аналитический набор данных для разработки может содержать сотни или даже тысячи переменных или метрик. Тем не менее он довольно мелок. Это означает, что во многих случаях разработка может производиться с помощью образца данных. Это делает ADS для разработки очень широким, но не очень глубоким.

Для скоринга и развертывания требуется производственный аналитический набор данных. Он включает только конкретные метрики, которые фактически содержались в итоговом решении. Как правило, большинство процессов потребуют только небольшой части метрик, изученных на стадии разработки. Основное отличие заключается в том, что процесс скоринга должен быть применен к каждому объекту, а не только к образцу. Каждый клиент, каждое местоположение, каждый товар должны быть оценены. Таким образом, производственный ADS будет не очень широким, но очень глубоким.

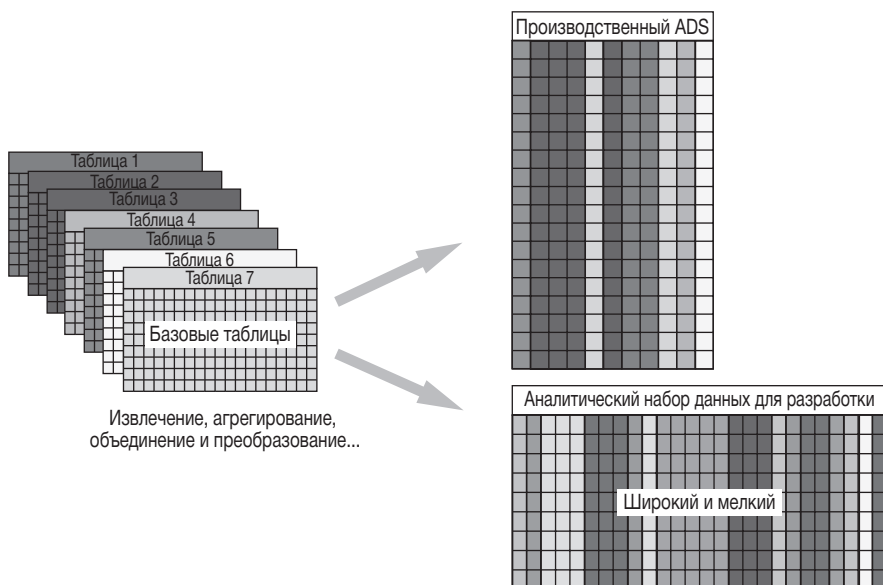


Рис. 5.5. Сравнение аналитического набора данных для разработки и производственного аналитического набора данных

Так, при разработке потребительской модели аналитик может исследовать 500 метрик для выборки 100 000 клиентов. Таким образом, аналитический набор данных для разработки широк, но неглубок. Когда придет время запуска процесса скоринга, потребуется, вероятно, только 12 метрик, но они будут нужны для всех 30 000 000 клиентов. Поэтому производственные ADS обычно узкие, но глубокие.

Традиционные аналитические наборы данных

В традиционной среде все аналитические наборы данных формируются за пределами базы данных (см. рис. 5.6). Каждый аналитик самостоятельно создает собственные аналитические наборы данных. Это означает, что сотни людей могут иметь свой собственный независимый взгляд на корпоративные данные. И ситуация ухудшается! Аналитический набор данных обычно создается с нуля для каждого конкретного проекта. Проблема состоит не только в том, что у каждого аналитика есть копия производственных данных. Каждый специалист часто создает новый ADS и, следовательно, новую копию данных для каждого проекта.

Процесс создания традиционного аналитического набора данных:
специальный ADS генерируется за пределами базы данных
для каждого проекта

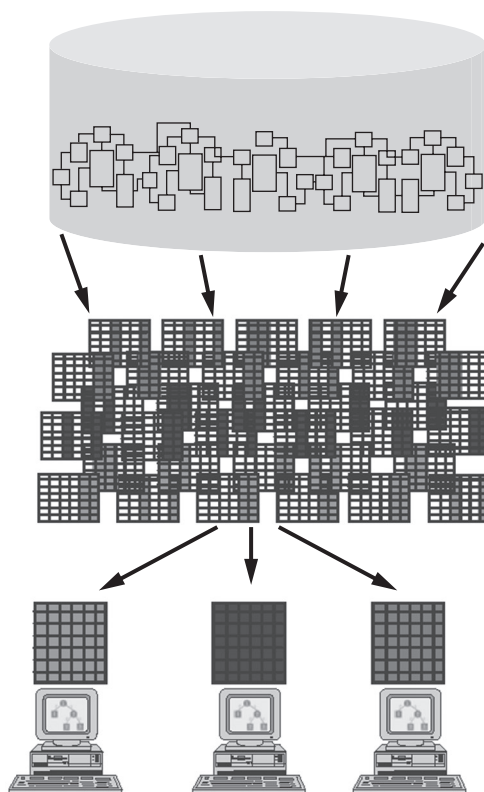


Рис. 5.6. Процесс создания традиционного аналитического набора данных

Бывает, что у компании, которая имеет определенный объем данных, в аналитической среде в итоге оказывается в 10 или в 20 раз больший объем данных. Когда организация переходит к современным масштабируемым процессам, она не желает переносить модель, подразумевающую наличие всех этих различных копий данных для каждого из пользователей. В данном случае требуется альтернативный метод, о котором речь впереди.

Одна из серьезных проблем, о которой люди не задумываются при работе с традиционными ADS, — риск несоответствия. Возможно, аналитик определил объем продаж как валовой объем продаж за вычетом

возвратов и скидок. В то же время человек в соседнем кабинете определил объем продаж как валовой объем продаж за вычетом скидок, но не возвратов. Нельзя сказать, что кто-либо из них абсолютно неправ — просто каждый использует несколько иное определение. Если они выполняют работу для одних и тех же деловых партнеров, то в их анализе и отчетах будут несоответствия. А это чревато проблемами.

Не менее серьезная проблема традиционного подхода к созданию аналитических наборов данных — необходимость повторения одних и тех же действий. Если аналитики снова и снова создают очень похожие наборы данных, то это означает, что они затрачивают не только дисковое пространство и системные ресурсы, но и свое время. Им приходится налаживать процессы создания ADS, запускать их и наблюдать за тем, чтобы они корректно выполнялись. Это увеличивает сроки сдачи и расходы на все проекты.

Несоответствие может нанести больший вред, чем избыточность данных

Хотя в процессе создания традиционного аналитического набора данных появляется большое количество избыточных данных, это не самая большая проблема. Часто из виду упускается тот факт, что разные аналитики могут использовать различные определения ключевых метрик. В результате появляются несоответствия. Они часто остаются нераспознанными или даже незамеченными.

Существует еще одна область, в которой энергия и ресурсы тратятся впустую. Создание ADS-процесса для проекта — еще не окончание работы. Когда процесс готов к производству, аналитику приходится перепроектировать его. Если, как это обычно бывает, производственная среда отличается от среды разработки, то для запуска проекта в производственной среде весь процесс приходится переписывать. Например, может потребоваться перевод кода с языка программирования на SQL или UDF. Это дорогостоящий процесс, часто связанный с ошибками. Некоторые компании тратят больше времени и денег на развертывание готовых аналитических процессов, чем на их создание!

Аналитический набор данных предприятия

Поговорим о способе упорядочения процесса создания аналитического набора данных с помощью аналитического набора данных предприятия, или EADS (enterprise analytic data set). EADS представляет собой

общий, допускающий многократное использование набор централизованных, стандартизированных аналитических наборов данных для применения в аналитических процессах.

Набор EADS сжимает сотни или тысячи переменных до небольшого количества таблиц и представлений, которые будут доступны для всех аналитиков, приложений и пользователей. В состав EADS может войти одна широкая таблица или несколько таблиц, соединенных в одну.

Различные аналитические процессы могут использовать один и тот же согласованный набор метрик EADS. EADS значительно упрощает доступ к данным, позволяя специалистам без дополнительных усилий извлекать множество показателей. Им больше не нужно самостоятельно извлекать показатели из необработанных таблиц, находящихся в третьей нормальной форме. EADS значительно сокращает время, необходимое на получение результатов; кроме того, его можно создать один раз, а использовать многократно (рис. 5.7).

Процесс создания аналитического набора данных предприятия:
централизованные таблицы и представления ADS
используются во многих проектах

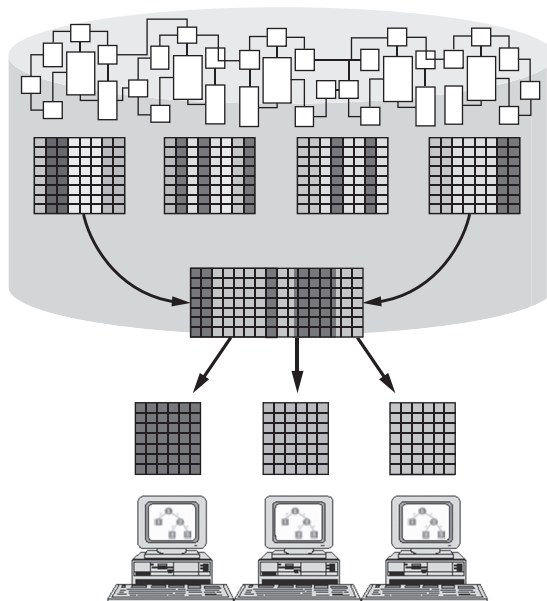


Рис. 5.7. Процесс создания аналитического набора данных предприятия

Одно из наиболее важных преимуществ EADS, о котором иногда забывают, — согласованность аналитических усилий. Большая согласованность в метриках, используемых в аналитических процессах организации, позволяет удостовериться, что эти показатели рассчитываются одинаково. При надлежащем использовании аналитический набор данных предприятия позволяет сократить время подготовки данных от 60–80% общего времени выполнения проекта до гораздо более низкого показателя. Целью может быть достижение показателя в 20–30%. Вот ключевые особенности аналитического набора данных предприятия:

- ▶ Стандартизированное представление данных для поддержания нескольких аналитических процессов.
- ▶ Способ, позволяющий значительно оптимизировать процесс подготовки данных.
- ▶ Способ обеспечения большей согласованности, точности и видимости для аналитических процессов.
- ▶ Способ обеспечения нового представления данных для приложений и пользователей за пределами пространства углубленной аналитики.
- ▶ Способ, который позволит специалистам тратить гораздо больше времени на анализ!

Когда следует создавать аналитический набор данных предприятия

EADS нужен, если вам часто приходится проводить анализ в конкретной области и вы ожидаете дальнейшего увеличения объема аналитической работы. Вы можете создать EADS для любого объекта, который анализируете: для потребителей, товаров, местоположения, сотрудников и поставщиков. Все, что анализируется на регулярной основе, может служить объектом EADS. Со временем EADS будет развиваться. При появлении нового источника больших данных в EADS могут быть добавлены дополнительные показатели для учета новой информации.

Создание аналитических наборов данных предприятия потребует временных и денежных затрат. Это не должно вас пугать. Со временем расходы окупятся благодаря экономии человеко-часов и аппаратных ресурсов. Концепция товарного моделирования, изложенная в главе 6, —

прекрасный пример анализа, провести который было бы невозможно без аналитического набора данных предприятия.

Для создания эффективного аналитического набора данных предприятия необходима кросс-функциональная команда. Специалистам в области бизнеса придется помочь определить показатели, которые они хотят изучить. Аналитикам нужно разработать логику генерирования этих показателей. ИТ-специалистам предстоит обеспечить поддержание структур и процессов аналитических наборов данных предприятия в среде, где они будут развернуты. Только если задействованы все три команды, можно максимально эффективно воспользоваться всеми преимуществами. В следующих разделах рассказывается, как происходит этот процесс.

Что входит в аналитический набор данных предприятия?

Проектирование аналитического набора данных предприятия — довольно простая задача. Процесс начинается с инвентаризации всех метрик, с которыми обычно работают аналитики. При наличии нескольких версий одного и того же показателя следует включить все. Например, некоторые аналитики используют валовой объем продаж за вычетом скидок и возвратов, а другие — валовой объем продаж только за вычетом скидок. Включите в EADS оба варианта так, чтобы он содержал валовой объем продаж за вычетом скидок и валовой объем продаж за вычетом как скидок, так и возвратов. Нет необходимости выбирать только один показатель! Если со временем появятся новые важные метрики, добавьте их. Аналитический набор данных предприятия должен постоянно развиваться. Поначалу EADS может содержать определенный набор метрик, который со временем будет дополняться.

Выбирайте всё!

В жизни редко предоставляется такой выбор, чтобы вы могли ответить: «Я возьму всё». Аналитический набор данных предприятия дает такую возможность. Используйте все варианты определения показателей, чтобы удовлетворить все потребности. Дополнительные усилия, связанные с добавлением дополнительных вариантов, практически несущественны. Оставьте споры о том, какая из метрик лучше подходит для той или иной задачи, на другой день. Вне зависимости от того, кто выигрывает в этом споре, ваши потребности будут учтены!

Важно понимать, что аналитический набор данных предприятия вовсе не должен предоставлять аналитикам 100% данных, которые могут потребоваться для каждого проекта. Этот набор может обеспечить порядка 90% данных; для конкретного же проекта могут быть нужны показатели, которые используются нечасто и поэтому отсутствуют в EADS. Это нормально.

Допустим, требуется провести глубокий анализ хитов продаж сезона отпусков. В наборе EADS, вероятно, окажутся только данные на уровне группы товаров. В этом случае анализ сосредоточен на небольшом наборе конкретных товаров. Необходимо вычислить показатели для этих товаров, чтобы дополнить ими уже присутствующие в EADS данные. Однако потребность в анализе этих отдельных товаров возникает нечасто, поэтому нет смысла добавлять их показатели в EADS.

Если аналитический набор данных предприятия содержит 80–90% необходимых метрик, аналитики могут сосредоточиться на вычислении дополнительных 10–20% показателей, а затем перейти к их анализу. Они также могут воспользоваться логикой, примененной при создании метрик EADS. Если со временем аналитики заметят, что часто добавляют одни и те же показатели, то эти показатели следует внести в EADS. Для добавления новых метрик придется разработать руководящие принципы.

Сравнение логической и физической структур

Уже упоминалось, что аналитический набор данных предприятия содержит одну строку для каждого объекта с десятками, сотнями или тысячами показателей. Если вы знакомы со «старыми» плоскими файлами, то примерно об этом и идет речь. Логическая и физическая структура EADS отличаются друг от друга (см. рис. 5.8).

Было бы логично представить EADS о клиенте в виде таблицы, содержащей данные о продажах, демографические данные и информацию, связанную с мероприятиями прямого маркетинга, однако данные о потребителе могут храниться иначе. Физическое представление EADS может включать одну таблицу, содержащую данные о продажах, одну таблицу с демографическими данными и одну таблицу с метриками, относящимися к мероприятиям прямого маркетинга.

Пользователям не следует об этом беспокоиться. Как только будут определены подходящие метрики, люди, управляющие базой данных,

смогут выбрать наилучший способ их хранения. Затем на основе физических таблиц можно добавить нужные пользователям представления.

Логическое представление EADS:

Таблица ADS, содержащая данные о потребителе

Потребитель	Общий объем продаж	Общий объем покупок	Домовладелец	Пол	Сервис автоматической рассылки писем	Разрешение на обращение по электронной почте
-------------	-----------------------	------------------------	--------------	-----	--	--

Возможное физическое представление EADS:

Данные о продажах

Потребитель	Общий объем продаж	Общий объем покупок
-------------	-----------------------	------------------------

Демографические данные

Потребитель	Домовладелец	Пол
-------------	--------------	-----

Прямой маркетинг

Потребитель	Сервис автоматической рассылки писем	Разрешение на обращение по электронной почте
-------------	--	--

Рис. 5.8. Сравнение логического и физического представления EADS

Обновление аналитического набора данных предприятия

Обновление аналитического набора данных предприятия — главная причина существования физически отдельных таблиц. Различные типы данных, например данные опросов, о продажах и демографические, могут требовать обновления с разной частотой: данные о продажах — ежедневного обновления; демографические — ежеквартально; данные опросов могут вообще никогда не обновляться. При проведении нового опроса данные загружаются в систему, и их уже не касаются.

Таким образом, легче поместить различные типы данных в разные физические таблицы, чтобы они могли обновляться независимо друг от друга. Это сэкономит системные ресурсы, поскольку в таблице не будет дополнительных метрик, когда в обновлении нуждаются лишь немногие. Кроме того, отдельные таблицы или представления облегчают аналитикам процесс извлечения конкретных типов данных, которые им требуются. Наконец, многие базы данных имеют ограничения на количество столбцов в одной таблице, поэтому для большого EADS может потребоваться несколько таблиц.

Обратите внимание: вне зависимости от того, как физически хранится EADS, по мере необходимости используются представления для

сбора различных фрагментов. Одно представление может содержать только показатели продаж и данные опросов, другое — только данные опросов и демографические показатели, а третье — все три типа данных. Со временем, если появится новый источник данных, например данные социальных медиа или веб-данные, основанные на этих данных метрики добавляют в аналитический набор данных предприятия. Можно определить и подходящий способ хранения новых данных, и обновленный набор представлений для их использования.

Сводные таблицы или представления?

Один из вариантов набора данных предприятия — набор сводных таблиц, которые обновляются с помощью запланированного процесса. У аналитического набора данных, основанного на таблицах, есть ряд преимуществ.

Во-первых, вы действительно можете вычислить показатели один раз и использовать их многократно. Общая нагрузка на систему, вызванная работой аналитиков, значительно уменьшится, поскольку вместо того, чтобы каждый специалист многократно запускал один и тот же тип процесса для объединения и агрегирования, этот процесс запускается один раз, а его результаты становятся общими.

Другое преимущество состоит в том, что большая часть передовых аналитических процессов подразумевает интенсивное использование исторических данных. Слегка устаревшие данные не окажут значительного влияния на результат. Возможно, организация обновляет данные о продажах в EADS каждую ночь или раз в неделю. Для большинства передовых аналитических проектов это нормально. Кумулятивные показатели также не окажутся сильно затронутыми. Например, когда для расчета показателя используются данные за предыдущий год, средний размер покупательской корзины не очень изменится, если не будет учтена сегодняшняя продажа.

Последнее преимущество заключается в том, что у аналитиков будет уходить мало времени на ожидание данных, поскольку они смогут обратиться к уже существующим таблицам EADS. Больше не нужно ждать, пока обработаются большие запросы. Специалисты смогут сразу приступить к анализу.

У EADS, основанного на таблицах, есть и недостатки. Первый: таблицы аналитического набора данных предприятия не будут содержать

самых последних данных. Второй заключается в том, что они будут использовать дисковое пространство системы, вероятно, довольно большое. Последний недостаток связан с необходимостью определения подходящего графика обновления для различных компонентов и налаживания соответствующих процессов.

Второй вариант набора данных предприятия — серия представлений, которые запускаются по требованию. У этого подхода есть несколько преимуществ.

Во-первых, аналитические наборы данных предприятия всегда будут полностью обновленными. Во-вторых, при необходимости проведения анализа в режиме реального времени у аналитиков не возникнет проблем, поскольку у них всегда будет доступ к самым свежим данным. Наконец, если в набор данных предприятия внести какие-либо изменения, то они окажутся доступными сразу же. Непосредственно после обновления представления следующий человек, который его запросит, получит доступ к новым данным.

Делайте то, что необходимо

Вам нужно решить, как часто следует обновлять аналитический набор данных предприятия. Вам также нужно решить, следует ли хранить EADS в виде физических таблиц, логических представлений или их сочетания. В каком направлении следует двигаться, помогут определить конкретные требования. В большинстве случаев используется комбинация таблиц и представлений.

У аналитических наборов данных предприятия, основанных на представлениях, также есть недостатки. Во-первых, нагрузка на систему не обязательно будет сильно снижена, поскольку, хотя аналитики используют одно и то же представление, процесс запускается каждый раз при обработке запроса. Далее, существует такое огромное преимущество, как согласованность и прозрачность вычислений. Последний недостаток связан с тем, что аналитикам придется дольше ждать, чтобы получить данные, поскольку они будут не вычисляться заранее, а генерироваться по необходимости.

Во многих случаях имеет смысл использовать в структуре EADS комбинацию таблиц и представлений. Одни данные будут обновляться постоянно, в то время как другие могут оказаться немного устаревшими. Обращайтесь с каждым конкретным источником данных наиболее подходящим способом. Решение об использовании таблицы или пред-

ставления должно быть основано на требованиях, связанных с производительностью и ограничениями пространства.

Используя таблицы, постарайтесь ограничить объем хранящихся данных. Не сохраняйте соотношения или другие аналогичные производные метрики — для вычисления таких показателей пользуйтесь представлениями поверх базовой таблицы. Например, если EADS содержит общий объем продаж и общее количество транзакций, нет смысла хранить и объем продаж на одну транзакцию. Создайте представление, которое делит объем продаж на количество транзакций. Вычисление этого показателя практически не требует дополнительных системных ресурсов и экономит много места.

Делитесь богатством!

После того как аналитический набор данных предприятия будет развернут, организации необходимо максимально эффективно его использовать, причем не только аналитиками. Нет причин, по которым структуры EADS не могут применяться в среде бизнес-аналитики и среде для создания отчетов. Зачем разрабатывать логику для вычисления метрик в среде для создания отчетов, если они уже содержатся в EADS?

Точно так же любое приложение, которое может извлечь пользу из содержимого EADS, должно использовать его. Один из распространенных примеров — CRM*, применяющий EADS для сегментации потребителей. В этом случае данные о клиентах, содержащиеся в EADS, становятся доступными для CRM-приложения. Основываясь на метриках, пользователи смогут выбрать клиентов, не вычисляя их с помощью CRM-инструмента. Примером может служить также приложение центра обработки вызовов, использующее EADS для предоставления данных о клиентах сотрудникам колл-центра. В этом случае, когда звонит клиент, сотрудник видит на экране различные данные о клиенте. Такие показатели, как недавно совершенные покупки, могут помочь сотруднику колл-центра решить, как лучше обработать данный вызов.

Дело в том, что в EADS содержится огромное количество информации. Он позволяет устранить дублирование усилий, значительно увеличить прозрачность и соответствие данных, обеспечить более высокую

* CRM — система управления взаимоотношениями с клиентами (англ. Customer Relationship Management).

скорость и масштабируемость. Не менее важным является тот факт, что EADS открывает прямой доступ к широкому спектру информации для других пользователей и приложений.

Встроенный скоринг

После создания аналитической песочницы и реализации аналитических наборов данных предприятия организации смогут более быстро и последовательно разрабатывать аналитические процессы и модели. Кроме того, аналитические процессы будут более масштабируемыми. Что дальше? Как ценность, создаваемая новыми процессами, может вывести организацию на следующий уровень? Один из способов — встроенный процесс скоринга, позволяющий использовать результаты анализа.

Встроенный скоринг подразумевает проведение скоринга в базе данных так, чтобы пользователи могли использовать созданные модели эффективным и масштабируемым образом. Успешная реализация встроенного скоринга будет включать в себя развертывание не только каждого отдельного процесса скоринга, но и процесса управления и отслеживания развернутых процессов. Обратите внимание: результатом проведения скоринга может являться показатель, полученный от прогнозной модели, или любой другой тип выходных аналитических данных.

Результатом аналитических процессов часто становится новый фрагмент информации. Примерами могут быть вероятность совершения покупки данным клиентом, точка оптимальной цены на товар или ожидаемое увеличение объема продаж в конкретном магазине в результате проведения акции. Когда разработанные аналитические процессы применяются к текущим данным, это и называется скорингом. Например, прежде чем решить, кому из потребителей стоит отправить электронное письмо, нужно определить вероятность отклика, основываясь на самых последних данных. Процесс обновления этих показателей вероятности представляет собой процесс скоринга, и он должен быть максимально автоматизирован и упрощен. Встраивание процессов скоринга в среду базы данных дает ряд преимуществ.

Во-первых, результаты пакетного скоринга будут доступны по требованию. Если набор оценок обновляется по графику, то пользователь в любой момент получит доступ к обновленным данным. Пакетное обновление также можно производить, когда это необходимо. Например,

организация будет обновлять оценки для клиентов, добавляемых в список рассылки, только в момент его создания.

Далее, встроенный скоринг позволяет производить оценку в режиме реального времени. Это особенно важно в случае с веб-предложениями. Посетитель, находящийся на сайте в данный момент, должен быть сейчас же оценен исходя из того, что о нем известно, включая то, что он только что сделал на сайте, чтобы предоставить ему подходящее предложение, когда он перейдет к следующей странице. То же верно и для ситуации, когда кто-то находится на линии с представителем колл-центра. Во время разговора представитель вводит любую новую информацию, которую он выяснил. Ввод этой информации может служить основанием для обновления оценки клиента, так что сотрудник колл-центра будет знать, что делать дальше.

Кроме того, встроенный скоринг избавит пользователей от излишней сложности. Отдельным пользователям и приложениям очень легко запросить оценку. Вся тяжелую работу выполняет система. В результате встроенный скоринг сделает оценки доступными для менее подкованных в техническом плане людей.

Последнее преимущество — наличие всех моделей в централизованном хранилище. Если инвентаризация созданных моделей и оценок осуществляется с помощью процесса управления моделями, отслеживать то, что было создано, становится еще легче. Аналитикам больше не придется держать под своим единоличным контролем создаваемые ими процессы скоринга. Напротив, эти процессы будут контролироваться централизованно и развертываться для более широкого использования.

Интеграция встроенного скоринга

После развертывания процедур скоринга генерируемые результаты легкодоступны и пользователям, и приложениям. Например, CRM-приложения могут использовать оценки склонности. Все, что необходимо сделать пользователю CRM-приложения, — это нажать соответствующую кнопку в программе, чтобы получить доступ к оценкам. Результаты скоринга также могут использоваться оперативными приложениями. Допустим, на основе показателей продаж модель предсказывает, что запасы товаров могут иссякнуть. При возникновении такой ситуации необходимо оповестить местного менеджера. Или, возможно, авиакомпания использует модели, определяющие вероятность задержки рейсов,

исходя из погодных условий. Эти оценки создаются для каждого рейса и передаются в приложение, используемое для отслеживания и управления этими задержками. Любой пользователь может получить прямой доступ к оценкам с помощью специального запроса.

Результаты проведения анализа должны использоваться для увеличения ценности

Для того чтобы воспользоваться преимуществами создаваемых аналитических процессов, организация должна применять их результаты. Не обеспечив легкий способ использования аналитики, организация не сможет в полном объеме реализовать весь ее потенциал. Встроенные процессы скоринга имеют решающее значение для обеспечения простоты применения их результатов широким кругом пользователей и приложений.

В главе 4 мы говорили о вариантах использования параллельных систем баз данных. Те же самые методы применимы для разработки встроенных процессов скоринга:

- SQL — родной язык базы данных — один из вариантов. Особенно это актуально для таких моделей, как деревья решений, линейная или логистическая регрессия. Даже написание кода вручную на языке SQL для таких моделей представляет собой достаточно простую задачу.
- Функции, определенные пользователем, на самом деле встраивают процедуру скоринга в базу данных как родную функцию.
- Язык разметки для прогнозного моделирования (PMML) представляет собой способ построения модели в одной системе и передачи информации о модели другой системе. Переданная информация позволяет принимающей системе автоматически сгенерировать код для процесса скоринга.
- Наконец, встроенные процессы позволяют аналитическим инструментам работать непосредственно в базе данных так, что не требуется перевода с языка аналитического инструмента на другие языки.

Обратитесь к главе 4, чтобы получить более подробную информацию о каждом из этих вариантов. Сейчас главное — понять, что к реализации встроенных процессов скоринга применимы все варианты.

Управление моделями и оценками

Для эффективного управления всеми аналитическими процессами предприятия необходимы четыре основных компонента: входные аналитические наборы данных, описание модели, проверка модели и отчетность, а также выходные данные модели (рис. 5.9). Существуют коммерческие инструменты для управления моделями и оценками, а также пользовательские решения, которые могут быть созданы для решения конкретных задач организации. Рассмотрим каждый компонент.



Рис. 5.9. Компоненты системы управления моделями и оценками

Входные аналитические наборы данных

Необходимо отслеживать детали, имеющие отношение к каждому аналитическому набору данных или аналитическому набору данных предприятия, которые используются в аналитических процессах, причем не только технические, касающиеся создания и хранения наборов данных. Нужно также отслеживать разнообразную информацию о них. Этот компонент системы управления моделями и оценками будет посвящен информации об аналитических наборах данных предприятия,

пользовательских наборах данных, созданных для конкретного процесса, или их сочетании. Отслеживаемые показатели включают следующее:

- Имя сценария SQL, хранимой процедуры, определенной пользователем функции, встроенного процесса, таблицы или представления, которые предоставят пользователю набор данных.
- Параметры, которые должны быть введены для запуска процесса аналитического набора данных. Например, пользователям может потребоваться указать диапазон дат или фильтр товаров.
- Выходная таблица(ы) и/или представление(я), которые будут созданы в результате запуска процесса, а также показатели, которые они будут содержать.
- Взаимосвязь между аналитическими наборами данных и созданными аналитическими процессами. Любой аналитический набор данных может использоваться одной или несколькими моделями или процессами. Конкретная модель или процесс также может потребовать более одного входного аналитического набора данных.

Описание модели

Необходимо отслеживать различные данные о каждой модели или процессе. Обратите внимание: моделью в этом случае может быть настоящая прогнозная модель или другой аналитический процесс, который должен запускаться на регулярной основе, например ранжирование клиентов в зависимости от объема продаж. Модель или процесс регистрируются в системе управления моделями в момент ее создания. Отслеживаемые данные включают следующее:

- Предполагаемый способ использования модели. Какую бизнес-проблему она решает? В каких ситуациях ее следует использовать?
- История модели. Когда она была создана? Кто ее создал? Какие изменения она претерпела?
- Статус модели. Находится ли она по-прежнему в разработке? Используется ли она? Выведена ли она из эксплуатации?

- ▶ Тип модели. Какой алгоритм использован? Какие методы применены?
- ▶ Скоринговая функция модели. Как называется SQL-сценарий, хранимая процедура, встроенный процесс или определенная пользователем функция, которая предоставляет пользователю результаты скоринга? Обратите внимание, что скоринговая функция предполагает доступность требуемых таблиц аналитического набора данных.
- ▶ Информация о входных переменных модели. Какие переменные из входного аналитического набора(ов) данных используются в модели или процессе? Конкретная модель или процесс может потребовать метрики только из одного ADS или из нескольких.

Проверка модели и отчетность

Как правило, для управления моделями и процессами необходимы отчеты, которые охватывают целый ряд тем и целей:

- ▶ Отчеты, которые сравнивают конкретный прогон процесса скоринга с базовыми показателями, использованными на стадии разработки.
- ▶ Сводная статистика, например график прироста, которая нуждается в изучении после каждого запуска процесса скоринга.
- ▶ Сравнение моделей или сводные данные о распределении переменных.

Отчеты могут создаваться автоматически при обновлении результатов скоринга или только по запросу. Такие отчеты часто используются на критической стадии мониторинга производительности модели. Со временем, по мере развития ситуации в бизнесе, модели будут деградировать. Отчеты помогают определить момент, когда модель следует пересмотреть.

Не теряйте контроль

Без согласованных усилий по отслеживанию моделей и аналитических процессов возникает риск, что модели будут использоваться неправильно или вообще не будут использоваться. Система управления моделями и оценками помогает снизить этот риск, а также гарантирует, что при обновлении одного процесса можно легко определить, какие из оставшихся процессов будут затронуты.

Выходные данные модели

Последние элементы, которые необходимо отслеживать, — это оценки, выходные данные процесса скоринга. Это фактические оценки, сгенерированные для каждого объекта, например потребителя, местоположения или товара:

- Каково значение оценки? Где она хранится? Каков идентификатор клиента, товара и т. д., для которого сгенерирована оценка?
- Временная отметка, соответствующая моменту создания оценки.
- При необходимости — созданные ранее оценки наряду с текущими. Одни организации в течение длительного времени хранят старые оценки, другие — нет. Следует выбрать тот вариант, который больше всего подходит для вашей организации.

Обзор главы

Самые важные уроки этой главы.

- Унаследованные процессы развертывания аналитических процессов и моделей не позволяют воспользоваться текущими условиями. Для укрощения больших данных очень важно обновить процессы, чтобы в полной мере воспользоваться доступными возможностями, связанными с масштабируемостью.
- Аналитикам нужны дополнительные разрешения. Аналитическая песочница представляет собой механизм, который дает им необходимую свободу и в то же время позволяет ИТ-отделу поддерживать баланс ресурсов.
- Песочница идеально подходит для исследования данных, разработки аналитических процессов и прототипирования. Ее не следует использовать для текущих или производственных процессов.
- Существует несколько типов среды песочницы, включая внутреннюю, внешнюю и гибридную. Каждый из них может быть дополнен средой MapReduce для облегчения процесса обработки источников больших данных.
- Аналитический набор данных (ADS) может содержать, к примеру, данные о потребителях, местоположениях, товарах и поставщиках.

- ▶ Не следует механически переносить традиционные, основанные на проектах подходы к работе с ADS на архитектуру, поддерживающую встроенную в базу данных аналитику. Вместо этого перейдите на использование аналитических наборов данных предприятия (EADS).
- ▶ EADS — набор предопределенных таблиц и представлений, которые предоставляют легкий доступ к сотням или тысячам часто используемых при анализе показателей.
- ▶ EADS улучшает производительность, устраняет избыточность, увеличивает прозрачность и согласованность аналитических инициатив.
- ▶ Предоставляйте доступ к EADS любым приложениям и пользователям, а не только аналитикам и аналитическим приложениям. EADS содержит важную информацию и должен находить более широкое применение.
- ▶ Встроенный скоринг использует как песочницы, так и структуры EADS для обеспечения процедур скоринга, к которым пользователи и приложения могут легко получить доступ.
- ▶ Скоринг может быть встроен с помощью кода SQL, определенной пользователем функции, встроенного процесса или языка PMML.
- ▶ Чтобы по-настоящему масштабировать использование моделей в организации, необходимо разработать систему управления моделями и оценками.
- ▶ Существует четыре основных компонента системы управления моделями и оценками: входные аналитические наборы данных, описание модели, проверка модели и отчетность и выходные данные модели.

Эволюция аналитических инструментов и методов

Можно ли построить дом, используя ручные инструменты и план десятилетней давности? Конечно, можно. Однако мало кто на это согласится, учитывая доступные сегодня электроинструменты и обновленные планы. Так же и аналитики могут продолжать разрабатывать аналитические процессы, используя только собственный код и традиционные методы. Однако если ознакомиться с имеющимися сегодня возможностями, мало кто захочет это делать. Сегодня отличный дом можно построить с меньшими физическими усилиями; то же верно и для аналитических процессов.

На протяжении многих лет аналитики использовали инструменты, которые позволяли им подготавливать данные для анализа, выполнять аналитические алгоритмы, а также оценивать результаты. Увеличение функциональности этих инструментов не вызывает удивления. В дополнение к гораздо более богатым пользовательским интерфейсам инструменты теперь дают возможность автоматизировать или оптимизировать решение распространенных задач. В результате аналитики могут больше времени посвящать анализу. Объединение новых инструментов и методов с более развитыми процессами и масштабируемостью (см. главы 4 и 5) предоставит организациям возможность укротить большие данные.

В этой главе пойдет речь о том, как аналитики изменили подход к построению аналитических процессов, чтобы более эффективно использовать новые доступные возможности инструментов и масштабируемости. Мы расскажем о групповом моделировании, моделировании

товаров и анализе текста, а также о том, как развивалось пространство аналитического инструмента и как эти усовершенствования будут продолжать изменять работу аналитиков. Мы поговорим об интерфейсах point-and-click («укажи и щелкни»), инструментах с открытым исходным кодом и инструментах визуализации данных.

Эволюция аналитических методов

Многие часто применяемые подходы к анализу и моделированию используются уже в течение многих лет. Некоторые из них, например линейная регрессия или деревья решения, эффективны и актуальны, но сильно упрощены. Раньше простота была продиктована жесткими ограничениями, обусловленными инструментами и масштабируемостью, однако сегодняшние возможности позволяют сделать гораздо больше.

До появления компьютеров было невозможно произвести многочисленные итерации модели или применить сложные методы. С увеличением масштаба технологий обработки данных увеличился масштаб инструментов и методов, используемых для их анализа. Сегодня можно множество раз применять разнообразные алгоритмы к большим наборам данных.

Нередко в результате увеличившейся масштабируемости специалистам просто приходится чаще прибегать к одним и тем же устоявшимся методам. Однако многие аналитики начинают применять новые методологии, которые позволяют лучше использовать усовершенствованные инструменты, процессы и возможности масштабируемости. Многие из этих новых методов были давно известны, но до последнего времени не применялись на практике. Это групповые методы, экспресс-моделирование и анализ текстовых данных.

Групповые методы (ensemble methods)

Групповые подходы концептуально достаточно просты. Вместо построения одной модели с помощью одной техники несколько моделей строятся с использованием нескольких техник. Как только результаты от всех моделей получены, они объединяются для определения итогового ответа. Для объединения различных результатов можно исполь-

зовать что угодно, от простого среднего показателя до гораздо более сложной формулы. Важно отметить, что групповые модели позволяют выйти за рамки выбора одной лучшей модели из набора. В данном случае объединяются результаты нескольких моделей для предоставления одного итогового ответа.

Мощь групповых моделей заключается в том, что различные методы имеют свои преимущества и недостатки. Например, некоторые типы клиентов могут получить плохую оценку при использовании одной техники, но очень хорошую — при использовании другой. Объединение данных, полученных от нескольких моделей, улучшает алгоритм скоринга в целом, если не буквально, для каждого оцененного клиента, товара или местоположения магазина.

Допустим, для оценки вероятности совершения покупки клиентом конкретного товара используются линейная регрессия, логистическая регрессия, дерево решений и нейронная сеть. Оценки, полученные от каждой модели, будут объединены в одну с помощью группового подхода. Часто это сочетание дает более надежное предсказание совершения покупки.

Групповым моделям посвящена отличная книга Джона Элдера и Джованни Сени «Групповые методы в интеллектуальном анализе данных» (Ensemble Methods in Data Mining by John Elder and Giovanni Seni¹¹). Групповые подходы получили распространение благодаря эволюции аналитических инструментов. Без наличия хорошего способа управления рабочим процессом и объединения результатов групповое моделирование представляет собой весьма громоздкий процесс. Представьте себе перспективу вручную запускать процесс для каждого из используемых методов. После завершения каждого процесса необходимо вручную объединить все результаты, чтобы оценить, как с задачей справился каждый из методов. Наконец, представьте, что необходимо решить, как объединить результаты в единый ответ. Сегодня аналитические инструменты могут сделать бóльшую часть или даже всю кропотливую работу за вас.

Мудрость толпы

Каждый отдельный способ моделирования имеет сильные и слабые стороны. Комбинируя различные результаты, мы получим единый ответ, который может быть лучше результата отдельных моделей. Это похоже на то, как усредненный ответ, основанный на предсказаниях множества людей, может оказаться близким к правильному. Это явление часто называют мудростью толпы.

Одна из причин растущей популярности групповых моделей заключается в простоте теории, лежащей в их основе. Мудрость толпы в повседневной жизни исследована довольно широко (см. книгу Джеймса Шуровьески «Мудрость толпы»¹²). Рынок предсказаний Iowa Electronic Market Университета штата Айова* в течение многих лет демонстрировал, что обоснованные предположения множества людей в среднем часто приближаются к правильному ответу. В самом деле, средний показатель может подойти к правильному ответу ближе, чем любой из ответов в отдельности.

Групповой метод использует концепции, которые изложены в книге «Мудрость толпы», применительно к аналитике. Множество моделей, делающих обоснованные предположения об исследуемых взаимосвязях, в среднем окажутся очень близки к правильному ответу. Может ли групповое моделирование решить все аналитические проблемы организации? Конечно, нет. Однако организациям следует добавить их в набор используемых методов.

Экспресс-модели (commodity models)

Одной из актуальных тенденций является использование так называемых экспресс-моделей. Мы определим экспресс-модель как модель, которая создается быстро и без особых попыток полностью реализовать весь ее предсказательный потенциал. Экспресс-модели могут создаваться, например, автоматически с помощью простой ступенчатой аналитической процедуры. Цель в данном случае не в построении наилучшей модели, а в быстром создании хоть какой-то модели, которая позволяет получить приемлемый результат.

При надлежащем использовании экспресс-модели весьма полезны в рамках организации. Раньше построение моделей требовало больших временных и денежных затрат. Аналитики тратили недели или месяцы только на сбор данных, а затем — на применение к этим данным созданных моделей, поэтому модели создавались редко и только для решения очень важных задач. Если бы вам предстояло разослать 30–40 миллионов писем с рекламным предложением, то в создание модели стоило

* Некоммерческий академический рынок, прогнозирующий результаты политических выборов. *Прим. ред.*

бы инвестировать. Однако если бы речь шла о предстоящей рассылке 30 000 предложений, касающихся недорогого товара, то инвестировать в создание модели было бы невыгодно.

Если аналитики используют современные среды, включая масштабируемые песочницы, а также современные процессы, в том числе аналитические наборы данных предприятия, то на построение модели уйдет намного меньше времени, чем раньше. Мы говорили об этом в главах 4 и 5. Чем более доступны эти стандартные переменные и чем бóльшие вычислительные мощности могут быть к ним применены, тем легче создавать модели.

Всегда помните, что легкость создания процесса не означает, что можно пренебречь необходимостью удостовериться в том, что этот процесс подходящий. Однако если им управляет хороший аналитик, вы добьетесь цели гораздо быстрее.

Иногда «достаточно хорошо» на самом деле означает «достаточно»!

Экспресс-модели призваны улучшить результаты там, где в противном случае вы вообще не использовали бы никаких моделей. Это более низкая планка по сравнению с той, которую всегда пытались преодолеть большинство моделей. Процесс создания экспресс-модели прекращается в момент достижения достаточно хорошего результата. Этот процесс хорошо подходит для решения задач малой важности или для ситуаций, когда требуется создать так много моделей, что их совершенствование не оправдано с прагматической точки зрения.

При оценке экспресс-модели основное внимание уделяется преимуществу, которое возникает благодаря ее использованию. Приложив дополнительные усилия, можно было бы многое улучшить. Однако если быстрая модель поможет в ситуации, в которой в противном случае модель бы не применялась, то она используется.

Приведем такую аналогию. Если у вас есть дом, то некоторые его части вы постараетесь сделать максимально удобными. Кухня, к примеру, требует особо тщательного подхода. В других случаях вам просто необходимо, чтобы работа была сделана. Возможно, что при переоборудовании гостевой ванной комнаты вы используете самые обычные материалы, поскольку в это помещение нет смысла вкладывать большие средства. Экспресс-модели помогают в подобных бизнес-ситуациях и имеют широкий спектр способов применения. Рассмотрим некоторые из них.

Способы применения экспресс-моделей

Экспресс-модели позволяют применить передовые аналитические методы к гораздо более широкому спектру задач и в более крупном масштабе в рамках организации, чем это возможно, когда аналитикам приходится вручную создавать модель за моделью.

Так, розничные торговцы часто создают модели «склонности к покупкам» для важных категорий товаров. Нет смысла создавать специальную модель для медленно развивающихся и реже продвигаемых категорий. Сети бакалейных магазинов следует создать модель для таких товаров, как моющие средства для ванны и газированные напитки. Создавать модель для товаров, пользующихся меньшим спросом, вроде крема для обуви или сардин, не имеет смысла.

Но что если возникает необходимость в продвижении менее важных товаров? Допустим, производитель сардин готов спонсировать проведение рекламной акции для своих товаров. Некоторые розничные торговцы сегодня имеют модели для всех своих многочисленных категорий товаров. Многие из них представляют собой экспресс-модели. Они создаются на случай, если понадобятся, и в этих ситуациях могут сформировать некоторую дополнительную ценность. Таким важным категориям, как газированные напитки или чистящие средства для ванной, по-прежнему уделяется особое внимание, и для них создаются отдельные более сложные модели. Тем не менее использование экспресс-моделей позволяет обеспечить менее важные категории товаров хотя бы простейшей моделью.

Сегодня благодаря аналитическим инструментам такие модели создаются легче. В них появились возможности автоматического выполнения алгоритмов с множеством комбинаций показателей и несколькими автоматизированными методами проверки. Это позволяет быстро создать довольно неплохую модель. Менее важные задачи потребуют другого подхода. В самом деле, нет ничего плохого в использовании достаточно хорошей модели вместо самой лучшей, когда ситуация этого требует.

Рассмотрим способ применения экспресс-моделей для прогнозирования. Представьте себе производителя, которому необходимо обеспечить как можно более надежные прогнозы относительно уровней спроса, например по кварталам, по товарам и по странам. Что если ему потребовалось бы спрогнозировать спрос в каждом магазине или точке

продаж на каждую неделю для каждого отдельного товара? На высококачественное прогнозирование просто не хватит человеко-часов. В таких случаях имеет смысл создавать автоматизированные достаточно хорошие прогнозы. Если прогнозы верхнего уровня точны, а совокупность прогнозов низкого уровня соответствует этой точности, то производитель останется доволен. В этом случае у него будут преимущества по сравнению с отсутствием каких-либо прогнозов.

Самое главное — убедиться в том, что вы используете процесс, который генерирует достаточно хорошие модели. Необходимо регулярно перепроверять процесс разработки экспресс-моделей и осмысленно оценивать их результаты. Не следует пускать процесс создания экспресс-моделей на самотек и позволять ему работать вообще без какого-либо вмешательства.

Анализ текста

Один из наиболее быстро развивающихся методов, используемых сегодня организациями, — анализ текста и других неструктурированных источников данных, к которым относится значительная часть больших данных. Анализ текста, как следует из названия, в качестве входных данных подразумевает некоторый текст. Он может представлять собой запись — электронное письмо, расшифровку диктофонной записи или даже отсканированный текст, преобразованный в электронную форму, например старые протоколы судебных заседаний. Причина роста популярности анализа текста — в богатстве новых источников текстовых данных.

В последние годы фиксируется все, начиная от электронной почты и комментариев в таких социальных сетях, как Facebook и Twitter, и заканчивая онлайн-запросами, текстовыми сообщениями и разговорами с сотрудниками колл-центров. Извлечение смысла из всех этих текстовых данных представляет собой непростую задачу. Существуют трудности, связанные с разбором, определением контекста и выявлением значимых закономерностей. Неструктурированных и текстовых данных у организаций становится больше, чем традиционных, структурированных данных. И эти типы данных нельзя игнорировать.

Текст — широко распространенный тип больших данных, и инструменты и методы его анализа прошли долгий путь развития. Сегодня существуют инструменты, которые помогают разобрать текст

на составляющие его слова и фразы, а затем определить значение этих слов и фраз. Популярные коммерческие инструменты анализа текста предлагают такие компании, как Attensity, Clarabridge, SAS и SPSS.

Разбив текст на компоненты, можно определить их настроение или значение и выявить существующие тенденции. Часто к сводным статистическим данным о разобранном тексте применяются модели. Например, сколько электронных писем конкретного клиента написаны в положительном или отрицательном тоне? Как часто данный клиент фокусируется на конкретной продуктовой линии в своих сообщениях? Это позволяет структурировать необработанную информацию. Такой способ разбора и структурирования текста часто называется извлечением информации.

Важно понять, что сами по себе неструктурированные данные не анализируются. Сначала они подвергаются обработке, в результате которой им придается некоторая структура. Затем производится анализ этих структурированных результатов. Вспомните сериалы, в которых детективы выявляют преступника. Берется отпечаток пальца, затем на него наносятся различные точки, которые соединяются между собой. Наконец, детективы находят совпадение и выявляют преступника. В данном случае производится сопоставление не исходного неструктурированного отпечатка, а созданной на основе его узора структурированной формы. Такой подход характерен для анализа источников больших неструктурированных данных.

Анализ неструктурированных данных

Как правило, сами по себе неструктурированные данные не подвергаются анализу. Сначала они подвергаются обработке, в результате которой им придается некоторая структура. Затем производится анализ структурированных результатов. Очень немногие аналитические процессы анализируют и делают выводы непосредственно на основе данных, находящихся в неструктурированной форме.

Применение контекста к текстовым данным представляет собой сложную задачу. Существуют определенные методы, однако этот процесс всегда предполагает долю творчества. Дело в том, что одни и те же слова могут иметь разные значения. Если я назову вас сумасшедшим, это будет воспринято как оскорбление. Однако если я скажу, что только что спустился по сумасшедшему горнолыжному склону, я имею в виду, что горнолыжный склон произвел потрясающее впечатление. Анализи-

ровать текст еще труднее, поскольку отдельные слова сами по себе часто не рассказывают всей истории и гораздо важнее то, как эти слова произносятся. Интонация может полностью изменить значение предложения.

Отличный пример приведен в табл. 6.1. Смысл всего предложения меняется, когда ударение перемещается. Если вы видите и слышите говорящего человека, то легко можете понять, что он имеет в виду. Когда у вас есть только текст, то понять его, используя лишь высказывание, невозможно. Предложения, которые окружают конкретное высказывание, помогают уяснить то, что имел в виду говорящий, однако переход на такой уровень анализа еще больше усложняет задачу. Вот почему анализ текста в течение некоторого времени будет оставаться сложной задачей.

Большинству организаций абсолютно необходимо начать использовать методы анализа текста. Анализ текста из метода, имеющего ограниченную область применения, превращается в технику, влияющую на широкий спектр отраслей и задач. Это один из примеров новых типов методик, которые необходимо развивать, чтобы обеспечить возможность обработки неструктурированных источников больших данных.

Табл. 6.1

Как акцент может изменить значение

Перенос акцента...	...изменяет значение
Я не говорил, что книга Билла — отстой	Но мой друг Боб сказал!
Я не говорил, что книга Билла — отстой	Как ты смеешь обвинять меня в этом?
Я не говорил , что книга Билла — отстой	Но я признаю, что написал это в электронном письме
Я не говорил, что книга Билла — отстой	Я сказал, что его блог — отстой!
Я не говорил, что книга Билла — отстой	Книга другого парня — отстой
Я не говорил, что книга Билла — отстой	Я просто сказал, что она не относится к числу моих любимых

Отслеживание появляющихся методов

Новые методы решения новых бизнес-задач появляются постоянно. Необходимо стремиться к тому, чтобы ваша организация использовала самые последние достижения. Если к вашему бизнесу применим новый

метод или подход, кому-то нужно будет в нем разобраться. Рассмотрим несколько методов, которые поначалу использовались редко, а потом распространились повсеместно. Эти примеры демонстрируют, насколько быстро редко применяемый метод может стать широко используемым.

Совместная фильтрация имеет цели, схожие с анализом близости. Оба подхода используются для того, чтобы выявить, в чем может быть заинтересован конкретный потребитель, исходя из интересов других, «похожих» клиентов. Совместная фильтрация используется сегодня на сайтах по всему миру и представляет собой довольно быстрый и надежный способ получения достойных рекомендаций. По сути, она обычно реализуется в форме экспресс-модели. Базовый подход легко развернуть и быстро получить рекомендации довольно хорошего качества. С развитием всемирной паутины совместная фильтрация получила довольно широкое распространение и влияние. Десять-пятнадцать лет назад этот метод не был так хорошо известен.

Алгоритм ранжирования страниц — это метод, лежащий в основе деятельности компании Google. Google использует его для определения наиболее релевантных ссылок, которые могут быть предоставлены пользователям при обработке поискового запроса. Все остальные поисковые системы располагают собственной версией алгоритма ранжирования страниц. Сегодня большинство отдельных сайтов имеют встроенный вариант этого алгоритма, используемого при осуществлении поиска на сайте. Эти методы были разработаны совсем недавно и не использовались до наступления эпохи интернета.

Большая часть населения никогда не слышала о совместной фильтрации или ранжировании страниц. Поколение назад большинство людей за всю свою жизнь могли ни разу не столкнуться с этими методами, однако в последние несколько лет они получили повсеместное распространение. Миллионы людей, путешествуя по всемирной паутине, используют эти методы анализа каждый день независимо от того, осознают они это или нет. В ближайшие годы широкое распространение получат другие методы, в настоящее время практически неизвестные. Каждая организация должна позаботиться о том, чтобы у нее были люди, которые отслеживали бы появление новых методов. О них можно узнать на конференциях по аналитическим технологиям, в специализированных журналах, статьях и блогах или от специалистов из других компаний.

Эволюция аналитических инструментов

Когда я занялся аналитикой в конце 1980-х годов, не существовало понятия «дружественный пользователю». Вся аналитическая работа выполнялась с помощью мейнфреймов. Для того чтобы провести анализ, приходилось не только непосредственно создавать программный код, но и использовать ужасный язык управления заданиями (JCL). Каждый, кто когда-либо сталкивался с JCL, знает, какая это головная боль!

Когда распространение получили серверы и ПК, они в основном представляли собой те же старые программные интерфейсы с новыми платформами. Графика и вывод данных в те времена находились в зачаточном состоянии. Первоначально графики генерировались с помощью текстовых знаков, из которых создавались столбиковые диаграммы, а для рисования сеток использовались тире. При выводе данных вы получали огромное количество текста с описанием того, что произошло.

Со временем были разработаны дополнительные графические интерфейсы, которые давали возможность вместо кодирования использовать среды point-and-click («укажи и щелкни»). Практически все коммерческие аналитические инструменты имели такие интерфейсы к концу 1990-х годов. С тех пор пользовательские интерфейсы были еще более усовершенствованы и теперь содержат более мощные графические средства, диаграммы потоков работ, а также приложения, сосредоточенные на конкретных точечных решениях. Диаграммы потоков работ — одна из самых полезных новых функций, поскольку они предоставляют аналитикам наглядную карту с отдельными этапами работы и связанными между собой задачами. Это позволяет визуально отслеживать все шаги процесса.

По мере развития инструментов продолжает развиваться и сфера их применения. В настоящее время существуют инструменты управления развертыванием аналитических процессов, управления и администрирования аналитических серверов и программного обеспечения, используемых аналитиками, а также инструменты для перевода кода с одного языка на другой. Кроме того, имеется ряд коммерческих аналитических пакетов. Хотя лидерами рынка остаются компании SAS и SPSS, существует множество других аналитических программ. Многие из них занимают определенную нишу, охватывая конкретные области. Кроме того, в настоящее время созданы аналитические инструменты с открытым исходным кодом. Об этом пойдет речь далее в главе.

Распространение графических пользовательских интерфейсов

До середины — конца 1990-х годов единственный вариант проведения статистического анализа подразумевал написание кода. Многие люди, особенно аналитики «старой школы», все еще любят писать код. Тем временем пользовательские интерфейсы становятся нормой, и аналитикам больше не нужно тратить много времени на кодирование. Графические пользовательские интерфейсы, доступные сегодня, позволяют генерировать большое количество кода «под капотом» от имени пользователей.

Часто можно слышать жаркие споры о том, используют ли «настоящие» профессионалы графический интерфейс или же они только пишут код. На самом деле ни у кого не должно возникать проблем с использованием графического интерфейса, пока он работает надежно и позволяет разрабатывать аналитические процессы в темпе, который равен или превышает темп ручного кодирования. Настоящие аналитики делают все возможное, чтобы выполнить работу настолько точно и эффективно, насколько это возможно. Кроме того, программное обеспечение сегодня предусматривает надежные решения, которые не только позволяют быстро генерировать код, но и помогают пользователям пройти через предопределенный процесс, направленный на решение конкретных проблем.

Дополнительное преимущество пользовательского интерфейса состоит в том, что автоматически сгенерированный код достаточно оптимизирован и свободен от ошибок. В этом заключается его отличие от ручного кодирования, при котором нередко опечатки, требуется отладка, а степень оптимизации производительности кода зависит от того, кто его написал. Ранние версии аналитических пользовательских интерфейсов были довольно громоздкими, и если человек умел хорошо кодировать, то ему быстрее было написать код, чем использовать интерфейс. Все изменилось с появлением новых пользовательских интерфейсов, которые эффективно автоматизируют генерирование большого количества кода. Это позволяет больше внимания уделять собственно анализу и необходимым методологиям и тратить меньше времени на кодирование.

С пользовательскими интерфейсами связана одна опасность, которая в то же время является одним из их ключевых преимуществ:

интерфейсы позволяют легко генерировать код. Звучит заманчиво, однако возможность быстро генерировать код также позволяет быстро генерировать и плохой код. Мы вернемся к этой теме в главе 8, когда будем говорить о том, что делает аналитика профессионалом. Если пользователь не профессионал, то с помощью пользовательского интерфейса он может случайно создать код, который будет делать совершенно не то, что задумано. Без понимания сгенерированного кода пользователь не в состоянии выявить такие ситуации, а это может привести к тому, что разработанные процессы будут некорректными или неточными.

Не надо быть старомодным

Многие пользовательские интерфейсы, доступные сегодня, действительно могут ускорить процесс генерирования кода, обеспечивая при этом отсутствие ошибок и оптимизацию. Специалистам стоит дать сегодняшним интерфейсам шанс. Результаты могут их удивить! Это особенно касается тех, кто на протяжении десятилетий занимался кодированием и сопротивляется любым другим способам. Инструменты сделают работу аналитиков более эффективной, освободив время для того, чтобы сосредоточиться на методах анализа вместо написания кода.

Пользователи графического интерфейса должны разбираться в кодировании и быть способными проверить код, сгенерированный за них инструментами, чтобы удостовериться в том, что созданный код делает именно то, что задумано. Часто при работе с пользовательским интерфейсом вы нажимаете несколько кнопок в ожидании определенного результата. Однако когда вы видите сгенерированный код, то понимаете, что результат отличается от того, что вы задумали. Современные пользовательские интерфейсы должны улучшать производительность, позволяя специалистам больше времени уделять анализу и меньше времени — кодированию. Однако инструменты не должны быть заменой знаниям, трудолюбию и усилиям.

Взрыв популярности точечных решений

В последнее десятилетие все чаще используются точечные аналитические решения, представляющие собой программные пакеты, направленные на решение очень узкого набора задач. Обычно они ориентированы на определенный набор связанных между собой бизнес-проблем и часто создаются на основе пакетов аналитических инструментов.

К точечным решениям относятся, например, приложения для оптимизации цен, выявления фактов мошенничества и прогнозирования спроса. Точечные решения, построенные на основе таких наборов инструментов, как SAS, используют некоторые общие функции базового пакета, однако пользовательский интерфейс настроен на решение конкретных задач. На разработку точечного решения может потребоваться много времени. Организациям следует подумать о приобретении такого решения вместо создания собственного. Это может сэкономить и деньги, и время.

Приложение для финансового учреждения, которое поможет обнаружить факты отмывания денег, например, должно предусматривать набор алгоритмов и бизнес-правил, которые находят подозрительные закономерности в движении средств. Интерфейс такого инструмента будет настроен на выявление подозрительных случаев и предоставление по мере необходимости дополнительной информации для помощи в процессе расследования. Такой инструмент может помочь организации быстро приступить к работе без необходимости разрабатывать множество процессов с нуля.

Аналитические точечные решения набирают популярность, поскольку позволяют различным отделам организации использовать более сложные аналитические методы в своих повседневных бизнес-процессах. Как правило, для того чтобы осуществить установку, конфигурирование и настройку параметров этих инструментов, необходим высокий уровень знаний. Однако их обслуживание и использование по силам менее подготовленным людям, что значительно расширяет пользовательскую базу точечных решений. Обратите внимание на то, что это не отменяет сказанного о людях, не использующих инструменты, если они не разбираются в кодировании. Точечные решения построены и сконфигурированы так, что пользователь совершает наиболее уместные действия.

Пользователи аналитических точечных решений, как правило, оказываются более продвинутыми по сравнению с рядовыми сотрудниками. Однако они не будут обладать такими же навыками, как профессиональные аналитики. Инструменты, сконфигурированные и настроенные специалистами, позволят автоматизировать решение многих задач, так что опытный пользователь сможет эффективно контролировать результаты работы инструмента и убедиться в том, что все работает нормально. Преимущество этого подхода заключается в более широком применении аналитики в рамках организации и в увеличении

масштаба. Ни у одной организации никогда не будет достаточно аналитиков для проведения всех необходимых видов анализа вручную. Аналитические точечные решения снимают часть этой нагрузки.

Используйте точечные решения

Аналитические точечные решения — отличный способ справиться с конкретными бизнес-проблемами. Такие инструменты позволяют подключить к аналитическому процессу больше людей. Использовать готовое коммерческое точечное решение намного быстрее, чем создавать собственное. Однако будьте готовы испытать потрясение, когда увидите цены на некоторые из доступных инструментов.

Серьезный недостаток точечных решений состоит в том, что они бывают весьма дорогостоящими. Некоторые точечные решения стоят порядка десяти миллионов долларов или больше, если речь идет об использовании инструмента в масштабах всего предприятия. Если ROI* оправдывает такие расходы, то это приемлемо. Однако типичная организация не может потратить достаточно денег, времени и усилий на реализацию многочисленных точечных решений, поэтому они нередко используются последовательно: по мере завершения реализации одного решения начинается реализация другого.

В ближайшие годы будут разработаны точечные решения применительно к некоторым аспектам анализа больших данных. Возможно, именно они потребуются организациям, чтобы начать такую работу. В процессе планирования своих действий следует изучить рынок, чтобы узнать о существующих возможностях.

История открытого программного обеспечения

Программные пакеты с открытым исходным кодом существуют довольно давно. Они доступны для всех и могут быть загружены бесплатно. Кроме того, сам код также доступен, поэтому при желании пользователи могут настраивать и добавлять функции в программное обеспечение.

Существуют примеры широко используемых и весьма успешных приложений с открытым исходным кодом: веб-браузер Firefox, операционная система Linux и веб-сервер Apache. Развитие интернета

* ROI (англ. return on investment — возврат инвестиций) — финансовый коэффициент (отношение суммы прибыли или убытков к сумме инвестиций), иллюстрирующий уровень доходности или убыточности бизнеса с учетом сделанных капиталовложений. ROI обычно выражается в процентах, реже — в виде дроби. *Прим. ред.*

способствовало росту активности, обусловленной открытым программным обеспечением. Если учесть все инновации, появившиеся в интернет пространстве, совершенно естественно, что им сопутствуют инновации в приложениях с открытым исходным кодом.

В настоящее время существуют самые разнообразные программные пакеты с открытым исходным кодом: базы данных, приложения для бизнес-аналитики и отчетности, инструменты для интеграции данных, офисные пакеты и т. д. В одних случаях, например в Linux и Apache, набор инструментов с открытым исходным кодом стал общепринятым вариантом, если не лидером, в своей области. Во многих других случаях (офисные средства) открытое программное обеспечение занимает конкретную нишу. Как правило, крупные и/или длительное время существующие корпорации медленнее принимают инструменты с открытым исходным кодом, чем новые сферы бизнеса или академическая среда.

Замечательная особенность инструментов с открытым исходным кодом состоит в том, что в повышение их функциональности свой вклад вносят тысячи людей. Обнаруженная ошибка достаточно быстро может быть исправлена многочисленными разработчиками-энтузиастами, которые работают в свободное время. Основные проекты с открытым исходным кодом поддерживаются формальными организациями. Есть организации, которые состоят полностью из добровольцев; на некоммерческих предприятиях, созданных для управления проектом, работают сотрудники, нанятые на полный рабочий день. За счет пожертвований эти организации могут платить сотрудникам зарплату, однако они не пытаются зарабатывать на самом программном обеспечении. Цель состоит в том, чтобы получить достаточное количество денег в виде пожертвований и, оплатив работу специалистов, гарантировать эффективное управление проектом. В будущем открытое программное обеспечение продолжит оказывать влияние, в том числе в сфере аналитики. Это подводит нас к проекту R.

Проект R для статистических расчетов

Открытое программное обеспечение пришло в мир передовой аналитики в виде «Проекта R для статистических расчетов», известного также как просто R. R — это бесплатно распространяемый аналитический пакет с открытым исходным кодом, который напрямую конкурирует

с коммерческими аналитическими инструментами, а также дополняет их. R — потомок S, одного из первых языков для статистического анализа, разработанного десятки лет назад. Свое название проект R, видимо, получил благодаря тому, что представлял собой обновление S, а также тому, что с буквы R начинаются имена его создателей (Роберт Джентлмен и Росс Айхэка)¹³.

Проект R быстро завоевал популярность и в настоящее время используется многочисленными профессиональными аналитиками. Это особенно верно в академической и исследовательской среде. Что касается корпоративной среды, то при наличии большой команды аналитиков по крайней мере некоторые из них так или иначе используют язык R.

Коммерческие инструменты по-прежнему доминируют, однако влияние R постепенно растет. Хотя количество его пользователей быстро увеличивается, на сегодняшний день он еще не так прочно прижился на крупных предприятиях, как в научных кругах. Язык R, как правило, используется для исследований и разработок, а не в крупномасштабных, критически важных для производства аналитических процессах. Со временем это может измениться, однако на момент написания данной книги дела обстоят именно так.

Язык R имеет широкий спектр возможностей. Он в большей степени объектно-ориентированный, чем многие другие наборы аналитических инструментов. Может быть связан с такими распространенными платформами программирования, как C++ и Java, что позволяет внедрять R-код в приложения. На самом деле коммерческие аналитические пакеты даже позволяют выполнять код, написанный на языке R, в рамках своих наборов инструментов. Это очень полезная функция. Более подробное описание этой темы выходит за рамки данной книги.

Возможно, самое большое преимущество языка R состоит в том, что при появлении нового метода моделирования или анализа кто-нибудь реализует его на этом языке. Функциональность R обновляется гораздо быстрее, чем функциональность коммерческих инструментов, и, если задуматься, так и должно быть. Поставщик коммерческого инструмента не будет спешить с интеграцией нового алгоритма, пока не убедится в том, что на его использование есть спрос. Убедившись в этом, поставщик добавляет этот алгоритм в свой график выхода, создает код и включает его в новую версию инструмента. На это могут уйти годы. В случае с R код алгоритма создается сразу, как только несколько человек сочтут его полезным.

То, что R распространяется бесплатно, для многих является несомненным преимуществом. Однако, как и в случае с любым другим проектом с открытым исходным кодом, существуют компании, которые предлагают собственные платные расширения и/или сервисы. Эти компании могут помочь вам с применением R, с разработкой процессов на языке R, а в некоторых случаях — предоставить вам расширения, улучшающие функциональность базового пакета. Недостаток бесплатного программного обеспечения — отсутствие поддержки. Вам приходится более или менее самостоятельно искать ответы на вопросы. Несмотря на существование многочисленного сообщества, нет ни одного конкретного ответственного человека или команды, к которым вы могли бы обратиться.

Вы используете R?

R — быстро развивающийся набор аналитических инструментов с открытым исходным кодом. За последние годы он сильно эволюционировал и получил широкое распространение. R имеет свои преимущества и недостатки и не подходит для каждой организации или каждой задачи. Однако в вашей организации он может сыграть определенную роль.

Один из главных недостатков языка R заключается в том, что программирование с его помощью — достаточно интенсивный процесс. Несмотря на существование графических интерфейсов, созданных на основе языка R, многие пользователи сегодня по-прежнему предпочитают писать код. Кроме того, R-интерфейсы гораздо менее развиты, чем аналогичные интерфейсы для коммерческих инструментов. Разумеется, со временем это может измениться.

Возможно, самым большим недостатком языка R является его плохая масштабируемость. В последнее время были сделаны некоторые улучшения, однако уровень масштабируемости R не соответствует уровню других коммерческих инструментов и баз данных. Компилятор R обрабатывает данные в оперативной памяти. Это означает, что он может работать только с наборами данных, размер которых соответствует объему доступной памяти компьютера. Даже у очень дорогого компьютера объем памяти гораздо меньше, чем требуется для работы с наборами данных предприятия, не говоря уже о больших данных. Если крупная организация хочет укротить большие данные, то R может стать частью решения, но не единственной, по крайней мере на сегодняшний день.

Все большее число инструментов, включая коммерческие аналитические пакеты, позволяют применять язык R. Станет ли он лидером, как Apache или Linux? Останется ли нишевым продуктом, как офисные пакеты с открытым исходным кодом? Только время покажет, какую роль будет играть R в сфере углубленной аналитики.

История визуализации данных

Визуализация данных так же стара, как и сами данные. В последнее время она превратилась в отдельную отрасль. Такие люди, как Эдвард Тафти*, зарабатывают на жизнь, обсуждая, исследуя и оценивая методы визуализации. Тафти написал множество книг, в том числе ставшую классической *Visual Display of Quantitative Information* («Визуальное отображение количественной информации»)¹⁴.

Изображенное Шарлем Жозефом Минаром уничтожение войск Наполеона во время похода на Москву в 1812 году считается одной из лучших визуализаций всех времен¹⁵. Посмотрев на это изображение, пройдя по ссылке в примечаниях к главе, вы сможете ясно представить себе, что произошло с этими войсками.

Визуализация в мире аналитики — это диаграммы, графики и таблицы, которые отображают данные. До появления компьютеров графики рисовали от руки. Компьютеры революционизировали и упростили методы создания визуализаций. Я помню свой первый цветной принтер для компьютера Radio Shack Color Computer. В нем в буквальном смысле были маленькие цветные шариковые ручки, рисующие на листе бумаги, похожем на широкую чековую ленту. Я мог создавать только некоторые очень примитивные столбиковые диаграммы с низким разрешением.

Раннее аналитическое программное обеспечение довольно умно использовало символы клавиатуры для создания графики, которая, возможно, и не была красивой, но очень хорошо позволяла донести смысл. Каждый столбик диаграммы мог состоять из ряда символов X (см. рис. 6.1); круговая диаграмма — из точек, запятых и тире, а рамку таблицы рисовали с помощью символов «» и «|».

Когда офисные приложения получили широкое распространение, практически у любого человека появилась возможность создавать

* Эдвард Тафти — профессор статистики Йельского университета, специалист по информационному дизайну. *Прим. ред.*

красочные диаграммы или графики с осями, подписями и легендами. Графические средства аналитических инструментов также сильно эволюционировали и вышли далеко за пределы создания графиков, состоящих из текстовых символов.

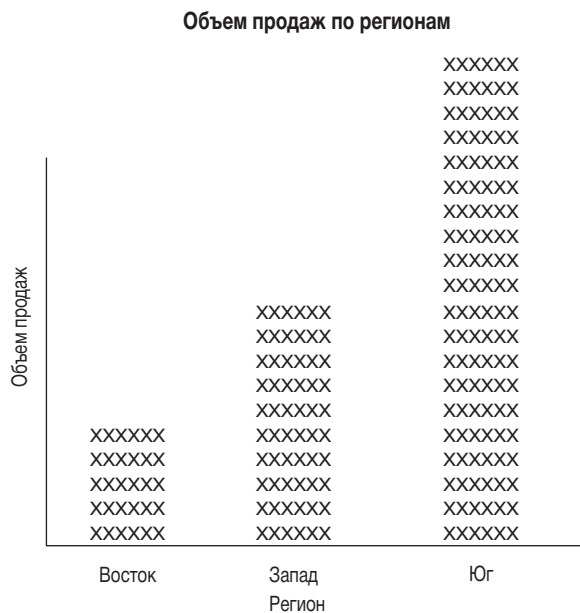


Рис. 6.1. Элементарная столбиковая диаграмма

Однако до недавнего времени визуализации в основном были статичными. Диаграмма в настольном приложении для создания презентаций или в электронной таблице оставалась статичной, пока не производилось ее обновление — как правило, вручную. Сегодня существуют средства визуализации, позволяющие взаимодействовать с графикой, исследуя и анализируя данные новыми и более эффективными способами.

Современные средства визуализации

Инструменты визуализации эволюционировали так сильно, что многие люди не осознают всех существующих возможностей. Такие инструменты, как Tableau, JMP, Advizor и Spotfire, помогают профессиональным аналитикам и бизнес-пользователям выйти за пределы графики,

которая просто иллюстрирует уже разработанную историю. Инструменты визуализации позволяют пользователю разработать новую историю, используя интерактивную визуальную парадигму.

Сегодняшние инструменты визуализации позволяют создать несколько вкладок с графиками и диаграммами, связанных с исходными данными. Еще более важно то, что вкладки, графики и диаграммы могут быть связаны друг с другом. Если пользователь щелкнет по столбику для северо-восточного региона, все остальные графики мгновенно скорректируются и будут отображать данные, относящиеся к этой области.

Эти новые инструменты можно представить как программное обеспечение для создания презентаций и электронных таблиц «на стероидах». Мало того что некоторые инструменты визуализации имеют такие же функции манипулирования данными, как и электронные таблицы, — они также обладают возможностями создания графиков, соперничающими с функциями приложений для создания презентаций или даже превосходящими их. Теперь добавьте к этому возможности подключения к большим базам данных, использования визуальных средств и углубленного изучения данных. В результате получается нечто очень мощное.

Основная предпосылка визуализации данных состоит в том, что очень трудно бывает разобраться в больших таблицах или наборах чисел и выявить тенденции. Гораздо легче увидеть тенденции, если используется подходящее визуальное представление. Некоторые визуализации, например графики, отображающие данные социальных сетей, передают информацию, которую было бы практически невозможно понять или описать без визуализации.

Только представьте себе попытку доходчиво объяснить человеку, как на карте расположены страны. Когда у вас перед глазами есть карта, вы точно знаете, как страны располагаются друг относительно друга. Очень сложно было бы придумать даже очень объемное объяснение, которое могло бы сравниться с картой по информативности и ясности.

Появилась новая идея аналитики с погружением (*immersive intelligence*), которая пока еще недоступна в коммерческих инструментах¹⁶. Она подразумевает использование возможностей трехмерной графики, онлайн-миров вроде *Second Life* и сложных визуальных инструментов (например, тех, что используются в области генетических исследований). Эти технологии применяются для интерактивного

представления данных. Будет ли возможна навигация по данным в интерактивной трехмерной среде для получения новых ценных сведений? Время покажет.

Не говорите — лучше один раз увидеть

Человеческий мозг очень хорошо интерпретирует визуальную информацию. Эффективная визуализация может помочь легко распознать закономерность или тенденцию. Глядя на традиционные электронные таблицы или отчеты, бывает сложно увидеть то, что вы ищете, и легко упустить важные взаимосвязи. Рисунок в виде эффективной визуализации данных может сказать больше, чем тысяча слов.

Визуализация помогает буквально увидеть новые идеи, которые иначе было бы невозможно обнаружить. Профессиональные аналитики в настоящее время используют эти инструменты для разработки аналитических процессов и исследования данных; некоторые специалисты прибегают к средствам визуализации исключительно для создания графики и презентаций. Эти инструменты гораздо быстрее и надежнее, чем традиционные графические. Кроме того, если во время презентации кто-то задает вопрос, можно, произведя анализ, получить ответ прямо в процессе представления, и не нужно обещать создать новый график и прислать его следующим утром. Любой организации, желающей укротить большие данные, следует рассмотреть вопрос о включении средств визуализации в свои наборы инструментов.

Важность визуализации для передовой аналитики

В главе 8 мы подробно расскажем, насколько важна для успешного анализа система коммуникации и доставки результатов. Профессиональному аналитику постоянно приходится объяснять сложные аналитические выводы деловым людям, не имеющим технической подготовки. Методы, которые позволяют делать это более эффективно, следует использовать. Визуализация данных относится именно к этой категории.

Зачем вдаваться во все детали логистической регрессии, если можно этого не делать? Включение всех оценок параметров, децилей и статистики оценки модели излишне, если простой график прироста скажет бизнес-спонсору все, что ему нужно знать. Подробности пригодятся в качестве резерва, однако бизнес-спонсоров не должны заботить технические детали. Они доверяют заботу о них своим аналитикам.

Немногие люди предпочтут увидеть длинный список бизнес-правил вместо наглядного дерева решений. Что если казино или розничному магазину необходимо определить самые оживленные области? Можно создать множество таблиц, разложить их на столе и попытаться в уме найти закономерности. А можно сделать тепловую карту пола казино или магазина, где цвет обозначает уровень активности. Ответ на поставленный вопрос будет очевиден сразу.

Важно впечатление, а не внешние эффекты

Важно, чтобы визуализация сразу делала идею очевидной. Слишком многие люди увязают в изощренной графике только потому, что могут себе это позволить. Простота — наилучший вариант. Эффектность или сложность должны быть оправданы.

Обратите внимание, что мы не говорим здесь о графике ради графики. Многие люди используют чрезмерную или слишком сложную графику только потому, что ее легко создать. Трехмерная столбиковая диаграмма не добавляет какой-либо аналитической ценности по сравнению с двумерной и может даже усложнить восприятие. Внимание должно быть сосредоточено на эффективной впечатляющей визуализации, которая помогает более четко проиллюстрировать идею. Красивая графика, которая не служит никакой цели, может отвлекать от основной мысли и приводить к путанице.

В одних случаях бывает достаточно простой таблицы. В других уместная визуализация может помочь аудитории гораздо лучше осмыслить идею. Вспомните пример с картой. Если аналитики понимают, как эффективно визуализировать данные и результаты, это поможет им стать более эффективными и успешными в своей работе. Средства визуализации только начинают оказывать влияние. В дальнейшем они будут использоваться все чаще в процессе анализа и представления его результатов.

Новые данные важнее новых инструментов и методов

Новые входные данные будут сильнее влиять на модель, чем новый инструмент или метод. Добавление новых данных в традиционный процесс обеспечит больший эффект, чем применение новых инструментов и методов к старым данным. Вот почему важно учиться работать с большими данными, а не просто обновлять методы работы с тем, что у вас есть.

Эта глава была посвящена достижениям в сфере инструментов и методов. Однако нужно помнить, что новые данные оказывают большее влияние на качество и эффективность аналитики, чем сами инструменты и методы. Так, например, наличие подробных веб-данных о потребителях, которых раньше не было, в большей мере будет способствовать повышению качества и эффективности модели склонности, чем достижению логистической регрессии или группового метода, используемого для построения модели. Новые инструменты помогают получить максимальную отдачу от новых источников данных, однако сами данные представляют собой более важный фактор. Вот почему для организаций очень важно использовать доступные им источники больших данных.

Обзор главы

Самые важные уроки этой главы.

- Групповые методы опираются на концепцию мудрости толпы. Объединение оценок, полученных при использовании нескольких подходов, может обеспечить лучший ответ, чем каждый отдельный подход сам по себе.
- Смысл использования экспресс-моделей заключается в быстром получении достаточно хорошей модели по возможности в автоматическом режиме. В данном случае достижение максимальной эффективности не самоцель.
- Экспресс-модели позволяют применить моделирование для решения менее важных задач, а также задач, требующих создания очень большого количества моделей.
- В эпоху больших данных анализ текста приобрел особую важность. Способы работы с текстовыми данными быстро развиваются и получают широкое применение.
- Трудности анализа текста заключаются в том, что слова сами по себе не рассказывают всей истории. Акцент и интонация имеют большое значение, однако в тексте отсутствуют данные о них.
- Пользовательские интерфейсы прошли большой путь развития и в данный момент включают в себя мощные графические средства, визуальные диаграммы потоков работ и узконаправленные точечные решения.

- ▶ Пользовательские интерфейсы должны применяться в качестве средств повышения производительности специалистов, которые разбираются в своем деле и могут убедиться, что «под капотом» инструменты делают именно то, что от них ожидается. Дружелюбный к пользователю интерфейс легко позволяет технически неподготовленным людям сделать что-нибудь неправильно.
- ▶ Аналитические точечные решения предназначены для эффективного решения узкого спектра аналитических задач, таких как выявление мошенничества или ценообразование. Такие инструменты приобретают все большую популярность.
- ▶ R — аналитический инструмент с открытым исходным кодом, который в последние годы получил широкое распространение. Преимущество R в скорости добавления новых алгоритмов, а недостаток — в невозможности на данный момент обеспечения масштабируемости до уровня предприятия.
- ▶ Увидеть закономерность гораздо легче, чем объяснить ее или выявить с помощью множества таблиц с данными. Современные средства визуализации позволяют обеспечить соединение с базой данных, создать интерактивные, связанные между собой графики и предоставляют гораздо больше вариантов визуализации, чем традиционные графические инструменты.
- ▶ Визуализация данных означает не использование изолированной графики, а отображение данных таким образом, который позволяет лучше понять доносимую идею.

ЧАСТЬ III

Укрощение больших данных: люди и подходы

Что такое хороший анализ?

Ведение статистики, написание отчета и применение алгоритма моделирования — лишь некоторые из множества шагов, необходимых для качественного анализа. Не существует «простой» кнопки, нажав на которую вы получите надежный результат. Если не проделать тщательной работы, это может привести к принятию неправильных решений и необходимости выполнения огромного объема дополнительной работы.

В этой главе начнем с уточнения некоторых определений, а затем покажем, чем отличаются отчетность или ведение статистики от анализа, а также качественный анализ от бесполезного.

Принципы работы, о которых пойдет речь, применимы не только к большим данным. Тем не менее, поскольку появление больших данных усложнило организациям жизнь, сегодня как никогда важно помнить об этих принципах. Ваша организация не сможет укротить большие данные только с помощью отчетов или некачественных средств анализа.

Анализ против отчетности

Многие организации ошибочно отождествляют анализ с отчетностью, поэтому следует разграничить эти понятия. Отчеты могут принести большую пользу, однако они имеют свои ограничения, и очень важно понимать, какие именно.

В конце концов, чтобы добиться успеха в процессе укрощения больших данных, организации потребуются и отчетность, и анализ, так же как в прошлом отчетность и анализ использовались для работы

с любым другим источником данных. Ключевой момент заключается в понимании разницы между отчетом и анализом; важно также понять, как они соотносятся друг с другом. Без этого ваша организация не сможет правильно их использовать.

Анализ производится мыслью

Анализ может привести к созданию отчетов, а отчеты — к анализу. Можно произвести анализ, основываясь исключительно на отчетах. Например, вы можете создать 10 отчетов, расположить их на столе, определить содержащиеся в них ключевые сведения и написать заключение: что вы обнаружили и что это означает. Это и будет анализом. Анализ создается мыслью аналитика, с помощью которой он оценивает значение данных или статистики для конкретного дела. Без интерпретации данные и статистика бесполезны.

Отчетность

Начнем с определения термина «отчетность». Среда для создания отчетов, как мы будем называть ее здесь, часто называется средой бизнес-аналитики (business intelligence, BI). В этой среде пользователи могут выбрать нужные отчеты, создать их и просмотреть результаты. Отчеты часто содержат таблицы, графики и диаграммы в любом сочетании. Вот ключевые характеристики отчетов:

- Отчет предоставляет пользователю данные, которые он запросил.
- Данные предоставляются в стандартизированном предопределенном формате.
- В создании отчета не задействуется никто, кроме пользователя, запросившего отчет через свой интерфейс. (Предполагается, что шаблон для отчета уже был создан и развернут.)
- В результате отчеты довольно негибки.

Уточним последний пункт. Сложные шаблоны отчетов создаются с различными подсказками и фильтрами. Такие отчеты могут содержать множество вариантов, однако в пределах этих предопределенных вариантов они довольно негибки. Среднестатистический пользователь, как правило, не в состоянии сформировать совершенно новый отчет или изменить работу подсказок и фильтров. Он может только использовать то, что уже есть.

Одна из распространенных ошибок состоит в принятии большого количества доступных отчетов за большой объем доступного анализа. Во многих организациях нередко можно услышать, как ИТ-специалист, отвечающий за среду бизнес-аналитики, говорит: «У нас есть высококлассная BI-среда, в которой доступно 500 отчетов, охватывающих все возможные аспекты бизнеса. У наших бизнесменов есть все, что им может потребоваться».

В то же время бизнес-пользователь говорит: «Я так расстроен! Мы потратили год или два на создание этой системы отчетности, а у меня по-прежнему нет того, что мне нужно». Если бизнес-пользователи и ИТ-специалисты окажутся в одной комнате, их разговор, скорее всего, начнется с жалоб бизнесменов о том, что у них нет того, что им нужно. ИТ-специалисты назовут их сумасшедшими, поскольку им доступно 500 отчетов. Это может перерасти в спор с взаимными обвинениями.

Разногласия коренятся в том, что пользователям, вероятно, действительно нужны эти 500 отчетов. Однако когда они их получают, смогут ли они извлечь из них необходимую информацию? Кроме того, два разных человека часто смотрят на вещи по-разному. Каждому из бизнес-пользователей может потребоваться одна дополнительная метрика в отчете или другая его организация. Среди 500 отчетов может отсутствовать именно тот, который необходим пользователю.

Отчеты: и количество не имеет значения!

Многие ИТ-организации создают отчеты, пытаются включить в них как можно больше тем. Это может быть обусловлено запросами бизнес-пользователей, которые охватывают все, что им может когда-либо потребоваться, а не то, что они на самом деле будут использовать. В результате пользователи теряются среди большого числа отчетов и не находят того, что им нужно. Следует сосредоточиться на предоставлении более ограниченного набора актуальных отчетов. Не попадайтесь в ловушку, думая, что выигрывает тот, кому доступно большее количество отчетов!

Лучше создать несколько отчетов, которые представляют собой именно то, что нужно конечным пользователям, чем всеобъемлющий набор из 500 отчетов. Важно не количество, а актуальность. Слишком часто внимание уделяется количеству отчетов (при этом считается, что «больше» означает «лучше»), а не актуальности. Даже наличие идеального набора отчетов для каждого бизнес-пользователя обеспечивает не анализ сам по себе, а доступность огромного количества данных для анализа.

Есть случаи, когда дальнейший анализ отчета не требуется. Предположим, у вас есть отчет об объемах продаж по каждому товару за неделю и вы хотите знать, достигли ли поставленных целей на прошлой неделе. Создав отчет, вы сразу получите ответ, и дальнейший анализ не потребуются. Именно так отчеты обеспечивают дополнительную ценность: помогают быстро и просто найти ответы на общие вопросы. Если все хорошо, то необходимость в дальнейшей работе отсутствует. Если результаты не соответствуют ожиданиям, нужно провести дальнейший анализ, чтобы выяснить причины.

Анализ

Теперь, когда мы разобрались с отчетностью, определим, что такое «анализ». Затем можно будет сравнить и сопоставить два этих понятия. Вот ключевые характеристики анализа:

- Анализ предоставляет ответы на вопросы, которые были заданы.
- Аналитический процесс включает любые шаги, необходимые для получения ответов на эти вопросы.
- Анализ подстраивается под конкретные решаемые вопросы.
- Анализ подразумевает человека, который руководит процессом.
- По своей природе процесс анализа гибок.

Анализ подразумевает следующее: «Я понял проблему. Я соберу все, что необходимо для ее решения». Это интерактивный процесс, когда человек разбирается в проблеме, находит данные, необходимые для получения ответа, анализирует их и интерпретирует результаты, чтобы предоставить рекомендации для дальнейших действий. Различия между анализом и отчетностью приведены в табл. 7.1.

Взаимодействие между отчетностью и анализом — распространенное и необходимое явление. В самом деле, одно повышает эффективность другого. Возьмем, к примеру, простейший отчет менеджера по продажам, содержащий данные о ежемесячных объемах продаж по регионам. Это очень простой отчет, который менеджер просматривает каждый день, чтобы определить, в правильном ли направлении развивается бизнес. Однажды менеджер обнаруживает в отчете нечто необычное и непонятное. Он идет к аналитикам и просит их разобраться

и выяснить, в чем дело. Его запрос, основанный на этом отчете, стал причиной проведения анализа.

С другой стороны, представьте себе аналитика, которому поручено исследовать эту проблему. Он выясняет, в чем заключаются ее причины, возвращается и показывает свои результаты менеджеру по продажам. Менеджер говорит, что собранные данные очень и очень полезны. Хотя они были сгенерированы для определения причин конкретной проблемы, он хочет получать эту же самую информацию на постоянной основе, даже если все идет по плану.

Что сейчас произошло? Анализ проблемы, проведенный сегодня, привел к появлению нового стандартного отчета. Аналитик автоматизирует свои действия, и они становятся стандартным отчетом.

Табл. 7.1

Сравнение анализа и отчетности

Отчетность...	Анализ...
Предоставляет данные	Предоставляет ответы
Предоставляет то, что запрашивалось	Предоставляет то, что нужно
Как правило, стандартизирована	Как правило, подстраивается под решение конкретной проблемы
Не задействует человека	Задействует человека
Характеризуется негибкостью	Чрезвычайно гибок

По мере того как ваша организация пытается освоить большие данные, следует иметь в виду, что хороший анализ можно предпринять, просто объединив имеющиеся данные новым способом для достижения новой цели. Это дает новое видение бизнеса. Как бы аналитики ни старались набивать себе цену, большая часть их работы связана с упрощенными вычислениями и подготовкой данных к анализу.

Ценность анализа — в ином восприятии данных

Цель анализа — не допустить ненужное усложнение проблемы. Иногда самый элементарный анализ дает все необходимые ответы. Простой взгляд на данные под другим углом помогает обнаружить ценные сведения. Если вы можете обойтись без лишних сложностей, так и сделайте. Порадуйтесь простому решению и переходите к следующей проблеме.

Очень часто получить ответ можно простым способом. Ценность заключается в том, чтобы подойти к делу иначе, а не в том, чтобы делать что-то сложное. Допустим, в объемах продаж розничной сети замечены некоторые аномалии. Можно создать сложную прогнозную модель, чтобы выяснить их ключевые факторы. А можно проверить цепи поставок. Не исключено, что доставка была произведена с задержкой или потребители остались дома из-за сложных погодных условий. Если удастся выявить такие причины, нет необходимости создавать сложную модель. Вы можете найти ответ на свой вопрос с помощью простого анализа и на этом остановиться.

Критерии хорошего анализа

Для того чтобы анализ имел значение, он должен быть хорошо сделан. Хороший анализ отвечает следующим критериям.

Актуальность

Хороший анализ руководствуется бизнес-потребностью. Он не производится только потому, что это интересно или весело. В случае с большими данными особенно легко втянуться в большое количество интересной, но бесполезной работы. Хороший анализ начинается с определения конкретной бизнес-задачи. Далее анализ производится исходя из того, что требуется для решения этой задачи. Каждый этап анализа должен руководствоваться потребностями решаемой проблемы.

Релевантность

Хороший анализ имеет отношение к бизнесу. Это означает, что бизнес-задача не должна быть произвольной; ее решение должно быть важным для бизнеса, а у бизнеса должна быть возможность ее решить. Не имеет смысла выяснять, насколько различные сегменты рынка чувствительны к цене товара, если его выпуск прекращен. Эти данные просто нерелевантны.

Объяснимость

Хороший анализ необходимо доходчиво объяснить тем, кому придется действовать, опираясь на его результаты. Легко увлечься формулами,

алгоритмами и статистикой. Технические детали, конечно, могут служить доказательством правильности проведенного анализа, однако результаты следует объяснять в терминах, понятных людям, принимающим решения, чтобы они могли воспользоваться его результатами.

Полезность

Хороший анализ является основанием для осуществления каких-либо действий. Он указывает конкретные шаги, которые можно предпринять, чтобы использовать результаты анализа для улучшения бизнеса. Нет смысла в анализе, который показывает, что перемещение нескольких магазинов на километр увеличит объем продаж, если у компании нет никакой возможности переместить магазины. Без возможности осуществить конкретные действия анализ бесполезен.

Своевременность

Хороший анализ предоставляется своевременно и доступен в момент принятия решения. Если решение необходимо принять на следующей неделе, получить ответ на свой вопрос в следующем месяце не имеет никакого смысла. Анализ может быть безупречным, но не быть завершенным в срок для принятия решения, к которому он имеет отношение. В этом случае выберите другую задачу и сосредоточьте усилия на ней. Запоздавший анализ попросту бесполезен.

Базовая аналитика против углубленной аналитики

В этой книге много говорится об углубленной аналитике. Чем же отличается «углубленная» аналитика от других видов аналитики? Для упрощения в качестве базовой аналитики рассмотрим неуглубленную аналитику. Базовая аналитика, как правило, связана с простыми вопросами и предоставлением простых ответов. Базовый аналитический процесс призван определить, что произошло, когда произошло и каковы последствия произошедшего.

Рассмотрим пример. Менеджеру по продукту необходимо оценить эффективность мероприятия по стимулированию сбыта, проведенного в прошлом месяце: удалось ли компании получить запланированное количество новых подписчиков? Базовый анализ может определить

количество новых подписчиков. Это говорит о том, что произошло. Анализ подписок по дням отвечает на вопрос, когда это произошло. Сколько денег принесли компании новые подписчики и как это соотносится с запланированными показателями? Это последствия.

Отметим, что все данные для проведения базового анализа могут быть обеспечены стандартизированными отчетами. Процесс анализа подразумевает необходимость просмотреть эти отчеты, сделать выводы и предложить возможные действия. В данном случае в процессе анализа показатели будут сравниваться с запланированными, чтобы определить, были ли достигнуты цели. На основании этого менеджер по продукту может сделать вывод о том, было ли проведенное мероприятие успешным.

Но дело в том, что такой базовый анализ не на все вопросы дает ответы. В частности, почему данное мероприятие привело к таким результатам и какие действия нужно предпринять в будущем?

Углубленная аналитика позволяет глубже разобраться в проблеме

Углубленная аналитика выходит за рамки ответа на вопросы о том, что произошло, когда это произошло и каковы последствия. Она пытается определить причины произошедшего и решить, что можно предпринять в будущем. Углубленная аналитика подразумевает ряд мероприятий, включая использование сложных *ad hoc* запросов SQL, прогностическое моделирование, интеллектуальный анализ данных, прогнозирование, оптимизацию и др.

Углубленная аналитика идет дальше, чем базовая, включая в себя процедуры от сложных *ad hoc* запросов SQL и прогнозирования до интеллектуального анализа данных и прогностического моделирования. Часто возникает вопрос о том, чем углубленная аналитика отличается от интеллектуального анализа данных, прогнозирования и прогностического моделирования. Дело в том, что все эти методы относятся к углубленной аналитике, а последняя также включает в себя и другие процессы, которые необязательно задействуют алгоритмы. Так, например, *ad hoc* запросы SQL — не повседневные запросы SQL, а очень сложные SQL-запросы, которые требуют комбинирования источников данных.

Причина, по которой *ad hoc* запросы SQL включены в определение углубленной аналитики, состоит в том, что ее основная цель — количественная оценка причин произошедших событий, предсказание возможного их повторения и определение того, как можно повлиять на эти

события в будущем. Иногда для того, чтобы получить ответы на эти вопросы, не требуется применять сложную модель.

В качестве примера рассмотрим компанию, которая проводит предварительное исследование веб-активности потребителей. Анализ должен определить, увеличивает ли просмотр информации о товаре в интернете вероятность совершения покупки. На разбор веб-данных и их объединение с другими сведениями о клиентах потребуется большое количество времени, поскольку веб-данные — явление новое. Отправной точкой может стать простой корреляционный анализ. На первом этапе нет необходимости создавать более сложную модель или процесс. А вот если между просмотрами страниц и объемами продаж обнаружится сильная корреляция, то компания может спокойно нацеливать свои маркетинговые мероприятия на людей, которые просмотрели страницу, но не совершили покупку. Возможно, позже компании потребуется провести более точную количественную оценку, однако в краткосрочной перспективе она будет использовать обнаруженную прибыльную закономерность.

Углубленная аналитика представляет собой важную часть общей аналитической стратегии организации. Она может помочь вывести организацию на новый уровень. Углубленная аналитика включает в себя очень сложные запросы SQL или манипулирование данными наряду с моделированием, прогнозированием, интеллектуальным анализом данных и другими подобными методами. Хотя в организации окажется совсем немного людей, обладающих навыками проведения углубленного анализа, они могут предоставить очень ценные сведения, которые в противном случае были бы недоступны.

Прислушивайтесь к своему анализу

Ни один анализ не может быть произведен хорошо, если к нему не подходить серьезно. Одна из часто встречающихся ловушек, о которых нужно знать, — так называемое «снятие сливок», то есть отбор результатов анализа на основании определенных критериев. Часто бывает, что в организации есть руководитель, работающий в течение очень долгого времени. Он привык принимать решения, основываясь в значительной степени на интуиции. Догадки такого руководителя, как правило, очень хороши. Впрочем, довольно трудно найти высокопоставленного

человека, который не мог бы принимать правильные решения чисто интуитивно. Такие руководители занимают свои позиции именно потому, что шестое чувство их не подводит. Однако цель анализа заключается не в замене интуиции руководителей, а в дополнении ее фактами.

Во многих случаях подобные руководители не желают позволять цифрам и данным указывать им, что делать. Некоторые из них запрашивают результаты анализа, чтобы посмотреть, поддерживают ли данные тот вариант действий, который они рассматривают. Если да, то он используется для оправдания решения и доказательства того, что это решение опирается на данные. Специалистам, которые произвели анализ, выражается глубокая благодарность. Звучит здорово, не правда ли?

Проблема возникает, когда результат анализа противоречит планам руководителя. Если компания и руководитель стремятся прислушиваться к аналитике и принимать решения, основываясь на фактах, то план следует пересмотреть. Однако иногда на результаты анализа закрываются глаза и план вводится в действие. О результатах анализа не упоминается, однако перечисляются все причины, по которым компании следует реализовать предложенный план.

Это и есть «снятие сливок». Только на основе результатов анализа можно достичь цели. Если вы решили их применять, то это необходимо делать повсеместно и последовательно. Использование результатов только тогда, когда они подтверждают ваши планы, не означает, что вы применяете аналитику для улучшения своего бизнеса. Это значит, что вы делаете то, что все равно бы сделали, и анализ для вас является лишь дополнительным подтверждением вашей правоты в тех случаях, когда он поддерживает вашу точку зрения. Поскольку результаты произведенного анализа не могут реально повлиять на принятие решений, проводить его не имеет смысла: он не дает никаких преимуществ.

Не поддерживайте «снятие сливок»!

«Снятие сливок» — это неправильный способ использования аналитики. Те, кто так делает, прислушиваются к результатам анализа, когда они поддерживают принятые решения, но игнорируют их, если они не оправдывают первоначальный план действий. Организация, утверждающая, что она использует аналитику для принятия решений, и в то же время поощряющая «снятие сливок», нечестна сама с собой. В таких условиях ничто не может измениться к лучшему. Будет лишь потрачено много дополнительного времени, денег и усилий на проведение анализа, который ничего не меняет.

Правильная постановка проблемы

Для того чтобы сделать хороший анализ, необходимо задать правильный вопрос, собрать нужные данные и разработать подходящий аналитический процесс, чтобы на этот вопрос ответить. Вероятно, самое важное различие между хорошим и плохим анализом заключается в правильной постановке вопроса. Необходимо начать с самого начала еще до запуска аналитического процесса.

Постановка вопроса подразумевает задание значимых вопросов и выдвижение критически важных предположений. Например, к чему должна привести новая инициатива — к обеспечению большей выручки или большей прибыли? Сделанный выбор имеет большое значение для последующего анализа и действий. Есть ли у вас все необходимые данные, или вам нужно собрать некоторую дополнительную информацию? Были ли рассмотрены альтернативы в плане разработки аналитического процесса для решения проблемы? Без постановки вопроса вся остальная работа не имеет смысла, поскольку результат будет абсолютно бесполезным.

Рассмотрим пример: команда консультантов создает для клиента модель сегментации. У клиента есть компонент «Бизнес—Бизнес» (B2B) и компонент «Бизнес—Потребитель» (B2C). Хотя консультанты знали о существовании компонента «Бизнес—Бизнес», он был незначительным и никогда не упоминался на встречах, на которых обсуждался проект.

Клиент отправил консультантам данные для проекта, однако у тех возникли затруднения из-за того, что некоторые клиенты отличались очень странным поведением. Консультанты сообщили клиенту о том, что обнаружили некие необычные закономерности, которые они не могут объяснить. Клиент немедленно сообщил им: «Это наши корпоративные клиенты». Клиент признал, что обсуждались только индивидуальные покупатели, в то время как предоставленные данные касаются обеих групп потребителей.

Включение данных о корпоративных клиентах — слишком важная информация, чтобы сообщать ее в последнюю очередь. Консультанты не учли корпоративных клиентов, и это привело к созданию неадекватных моделей. В конце концов консультанты создали две модели: для корпоративных клиентов и для индивидуальных потребителей. Две эти группы необходимо было разделить из-за значительной разницы

в поведении. Для правильной постановки проблемы необходимо было либо сосредоточиться только на одном типе клиентов, либо создать модель для каждого типа.

Самое главное — постановка проблемы

То, как вы сформулируете проблему и разработаете аналитический процесс, важнее, чем все, что вы будете делать дальше. Анализ не может быть точным и полезным, если проблема сформулирована некорректно и разработан неподходящий аналитический процесс. Уделите должное внимание процессам формулирования и разработки, чтобы гарантировать правильную реализацию того и другого. В противном случае хорошего результата добиться не удастся.

Хороший анализ начинается с правильной постановки проблемы. Это означает правильно оценить данные, разработать план проведения анализа и принять во внимание различные технические и практические аспекты, имеющие отношение к делу. Можно утверждать, что постановка проблемы — важнейший шаг анализа, поскольку его неправильное выполнение приводит к неправильным дальнейшим действиям.

Статистическая значимость и важность для бизнеса

Профессиональные аналитики уделяют много внимания статистической значимости, и это хорошо. Однако статистическая значимость — лишь один из аспектов хорошего анализа. Проверка статистической значимости подразумевает выдвижение ряда предположений и определение вероятности того, что полученные результаты имели бы место в случае правильности выдвинутых предположений.

Например, если предполагается, что монета симметрична, то количество случаев выпадения орла и решки будет одинаковым. При подбрасывании симметричной монеты шансы выпадения решки 10 раз подряд очень малы. Если это случилось, то существуют только два возможных объяснения. Первое заключается в том, что это была полоса везения, что встречается один раз из 1024 попыток. Второе — в том, что монета несимметрична. Проверка статистической значимости, касающаяся выпадения решки 10 раз подряд, показала бы, что вероятность того, что монета несимметрична, равна 99,9%, поскольку симметричная монета

позволяет получить такой результат лишь в 0,1% случаев. Такие расчеты представляют собой суть статистической значимости.

Необходимо различать статистическую значимость и важность для бизнеса. Это не одно и то же.

Статистическая значимость

Статистическая значимость часто используется для выявления средних значений и процентов, а также для определения оценок параметров статистических моделей. Проверка статистической значимости поможет убедиться в том, что данные не вводят вас в заблуждение. Она с математической точки зрения покажет, достаточно ли значимо различие. Бывает, что различия, которые кажутся существенными, не являются таковыми, а бывает и так, что значимыми оказываются небольшие различия. Статистическая проверка позволит убедиться в правильности сделанных выводов.

На основе тестирования создана целая дисциплина. В деловом мире она известна как подход «тестируй и изучай» (test and learn), включающий основные экспериментальные концепции, которые преподаются на курсах статистики в колледже. В среде «тестируй и изучай» эксперимент устроен так, что можно измерить эффекты использования одного или нескольких вариантов и определить, какой из них будет работать лучше всего.

Предприятия должны удостовериться, что они используют правильный подход и не гонятся за «очевидным» ответом. Один из моих любимых примеров, противоречащих интуиции, — задача, которую задают студентам в магистратуре. Посмотрите на табл. 7.2. Два игрока в бейсбол играли вместе в течение пяти сезонов. Согласно таблице в каждом сезоне среднее количество отбивания на бите у Джо было выше, чем у Тома. Возникает простой вопрос: «Кто из двух игроков имеет больший средний показатель за все пять сезонов?» Задумайтесь на минуту и запомните свой ответ.

Правильный ответ может удивить вас: мы не знаем, кто имеет больший средний показатель! Информации, представленной в табл. 7.2, недостаточно, чтобы ответить на этот вопрос. Почему? Если бы мы знали, что у Джо и Тома было одинаковое количество выходов на биту в каждом сезоне, то ответ был бы столь же простым, как мог показаться на первый взгляд. Победителем был бы Джо. Но что если у них было

разное количество выходов на биту? Что если в тот сезон, когда у обоих игроков было наилучшее среднее количество отбивания, Джо получил травму и за несколько месяцев выходил на биту намного реже, чем Том? Что если Том получил травму в сезон с самыми низкими показателями и, таким образом, Джо выходил на биту намного чаще? Получается, что у Тома может быть более высокий совокупный средний показатель, чем у Джо, хотя в каждом сезоне показатель Тома был ниже! Так бывает нечасто, но все же бывает!

Табл. 7.2

Среднее количество отбивания на бите по сезонам

Сезон	Том	Джо	Победитель
1	252	255	Джо
2	259	266	Джо
3	237	241	Джо
4	253	255	Джо
5	256	257	Джо

Никогда не «срезайте углы»

Если у вас неполная информация, легко прийти к неправильным выводам. Никогда не выбирайте легкий путь и не считайте результаты настолько убедительными, что нет необходимости в формальном доказательстве их статистической значимости. Всегда старайтесь убедиться в наличии всех необходимых данных и проверьте эти данные, прежде чем делать выводы.

Не зная количества выходов на биту, невозможно определить лучшего игрока. Посмотрите на табл. 7.3, чтобы понять, как Том может оказаться победителем по результатам пяти сезонов. В данном случае *t*-тест показал, что разница между средними показателями Тома и Джо не является статистически значимой. Таким образом, вместо очевидного ответа, что Джо имеет лучший показатель по сравнению с Томом, мы выясняем, что победитель — Том. Но и здесь не все так просто! Хотя Том победил, разница не является статистически значимой. С точки зрения статистики у них ничья. Ответ на этот вопрос имеет больше нюансов, чем кажется на первый взгляд.

Табл. 7.3

Полное сравнение средних показателей

Год	Том: средний показатель	Том: выходы на биту	Том: количество ударов	Джо: средний показатель	Джо: выходы на биту	Джо: количество ударов	Победитель
1	.252	123	31	.255	341	87	Джо
2	.259	355	92	.256	109	29	
3	.237	139	33	.241	377	91	
4	.253	304	77	.255	294	75	
5	.256	363	93	.257	206	53	
Общий показатель	.254	1284	326	.252	1327	335	Том*

* Том победил, но разница в показателях не является статистически значимой. С точки зрения статистики у них ничья.

Большинство людей посмотрят табл. 7.2 и не станут утруждать себя лишними размышлениями. Они выберут очевидный ответ: у Джо лучший общий средний показатель. Никогда так не делайте! Всегда проверяйте свои выводы.

И последний момент, касающийся статистической значимости. Большинство людей будут удовлетворены, если получают в результате эксперимента 95%- или 99%-ную вероятность. Следует, однако, иметь в виду, что, когда вы на 95% уверены в том, что правы, существует еще 5%-ная вероятность того, что вы ошибаетесь. Это означает, что один раз из каждых 20 случаев проведения подобного эксперимента вы можете оказаться неправы, соглашаясь с результатами.

Убедитесь, что уровень доверия соответствует уровню риска, который вы можете себе позволить. Например, если в случае выбора неправильного ответа компании грозит полное банкротство, то 95%-ная вероятность не кажется таким уж хорошим показателем. Вероятно, следует стремиться к показателю 99,9% или выше.

В случае многочисленных повторений одного и того же действия шансы ошибиться по крайней мере один раз значительно повышаются. Вы должны быть готовы принять эти ошибки или установить очень высокий уровень доверия, чтобы обеспечить очень низкий уровень риска. Для клинических испытаний новых лекарственных препаратов

используются очень высокие показатели, поскольку последствия выпуска плохого лекарства велики, вплоть до смерти. Уровень доверия для принятия решения о том, какое из двух изображений компании стоит поместить в верхней части веб-страницы на остаток дня, может быть значительно ниже.

Важность для бизнеса

Рассмотрев тему статистической значимости, мы поняли необходимость сбора полной информации и проведения правильных тестов. Кроме того, убедились, что никто не может быть на 100% уверенным в правильности принятого решения. История на этом не заканчивается. Финальный шаг — оценка важности статистически значимого вывода для бизнеса.

Предположим, в ходе анализа был сделан статистически значимый вывод. Однако существует ряд других не менее важных, а может быть, даже более важных вопросов. Есть статистически значимый результат, и это замечательно. Но представляет ли он важность для бизнеса? Как организация может использовать этот результат для принятия решения о дальнейших действиях? Был выявлен реальный эффект, но достаточно ли он сильный, чтобы привести к значительным последствиям?

Всегда проверяйте результаты в бизнес-контексте. Допустим, существует 99%-ная вероятность того, что данное изменение в предложении может привести к 10%-ному повышению уровня отклика. Это хорошо. Но что если базовым является первоначальное предложение, а тестируемое изменение — это бонусное предложение, которое стоит в два раза дороже? В этом случае 10%-ное увеличение количества откликов может не покрыть дополнительные расходы. Тот факт, что доля откликнувшихся может быть существенно увеличена, на самом деле не имеет значения с точки зрения бизнеса.

Выйдите за рамки тестирования статистической значимости и постарайтесь учесть более полную картину. Какие затраты связаны с введением рекомендуемых изменений? Какой дополнительный доход может быть получен со временем? Соответствует ли новый подход общей корпоративной стратегии? Достаточно ли у вас людей и времени для внесения необходимых изменений? Статистическая значимость — критически важный показатель, но она имеет значение только в том случае, если рассматриваемое изменение важно с точки зрения бизнеса.

Дополнительную ценность обеспечивает хороший анализ

Очень важно понимать разницу между статистической значимостью и важностью для бизнеса, особенно в условиях надвигающегося вала больших данных. Профессиональные аналитики всегда найдут в больших данных действительно интересные сведения. Они помешаны на числах и поэтому могут воскликнуть: «Вот это да! Здорово!» Однако важно задать вопрос, насколько ценны и релевантны для бизнеса эти сведения. Это неотъемлемая часть анализа. Если ответ «нет», то это просто шумиха.

Выборка против популяции

Прежде использование выборок считалось обычной практикой. Необходимо только было обеспечить, чтобы размер выборки был достаточным для решения поставленной проблемы. В случае с большими данными наличие достаточного объема данных для выборки, конечно, не проблема. Используя сегодняшние масштабируемые системы, можно работать с целой популяцией. Больше нет необходимости в отборе 10% клиентов в связи с невозможностью обработки большего объема информации. В некоторых областях, например клинических испытаниях, небольшие размеры выборки все еще могут создавать трудности. Сегодня эти области скорее исключение, чем правило. Тем не менее по-прежнему важно определять, когда в процессе анализа следует использовать выборки. При этом их следует создавать правильно.

В следующий раз, когда будете читать газету, обратите внимание на содержащиеся в ней результаты исследования. Под ними будет указан предел погрешности; как правило, это плюс-минус 3–5%. Вы также увидите размер выборки, который обычно составляет от 800 до 1200 человек. Предел погрешности и размер выборки будут примерно одинаковым вне зависимости от вопроса, темы и размера популяции, из которой была сделана выборка. Все, что необходимо для того, чтобы оказаться в пределах нескольких процентных пунктов, — это получить около 1000 ответов.

Чем больше размер выборки, тем меньше погрешность и выше вероятность того, что «правильный» ответ очень близок к тому, который был найден на основе выборки. Большие данные обеспечивают такие размеры выборок, на основе которых можно получить результаты с очень высокими уровнями статистической значимости. Однако различия могут быть чрезвычайно малыми и незначительными с точки зрения бизнеса.

Допустим, производится исследование сотен миллионов веб-сессий, чтобы определить, сколько людей перешли по ссылке А и по ссылке В. При этом установлено, что 2,5235% людей щелкнули по ссылке А, а 2,5237% — по ссылке В. Эта разница в 0,0002% может быть статистически значимой, если выборка достаточно велика. Тем не менее разница несущественна. Она не удовлетворяет критерию важности для бизнеса или критерию релевантности. Как гласит старое правило статистики, «разница должна быть значимой».

Раньше аналитики переживали из-за размера выборки. Беспокойство было вызвано тем, что при использовании маленькой выборки погрешность бывает слишком большой. Когда выборка чрезмерно мала, разница должна быть относительно большой, чтобы считаться статистически значимой. В таких условиях проведение анализа часто бывает бессмысленным. В настоящее время необходимо убедиться, что выборка не слишком велика. Понятие «слишком большой» выборки кажется странным. Но это следует иметь в виду.

Если для решения конкретной проблемы требуется использовать выборку, состоящую из 200 000 клиентов, чтобы обеспечить необходимую степень точности, то обработка выборки в 2 000 000 только потому, что это возможно технически, будет пустой тратой времени и ресурсов. Выборка должна иметь такой размер, который позволит найти статистически значимую разницу, имеющую важность для бизнеса. Если для принятия мер необходимо выявить разницу в 1%, то выберите такой размер выборки, при использовании которого разница в 1% будет иметь статистическую значимость. При использовании слишком большой выборки статистическую значимость может иметь разница, равная доле процента. При этом производится лишняя обработка, не имеющая никакой практической пользы. Убедитесь, что вы используете достаточно большую выборку, размер которой, однако, не слишком превышает минимально необходимый. Укрощение больших данных потребует уменьшения объема до необходимых размеров.

В некоторых случаях может потребоваться использование 100% данных. Один из наиболее распространенных примеров — необходимость создания списка «Лучших N» на основе некоторого критерия. Например, требуется выявить 100 клиентов с самыми большими расходами. По определению любая случайная выборка не может включать всех лучших клиентов, в ней может оказаться лишь случайное их подмножество. Чтобы определить сто лучших клиентов, необходимо учесть их всех.

Как и прежде, сама проблема определяет необходимость в применении выборки и ее размер. По возможности старайтесь использовать выборки эффективно.

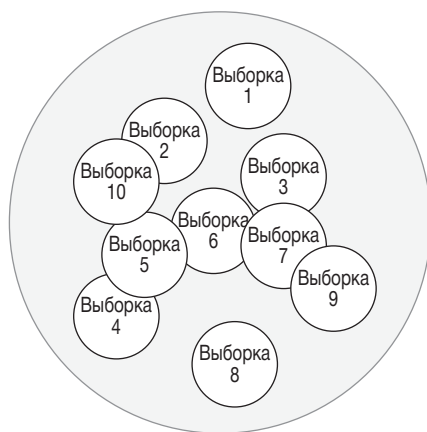
Существует и такое распространенное заблуждение: одна и та же выборка подходит для различных задач. Например, отделу маркетинга требуется выборка размером в 10% от общего числа клиентов. Маркетологи создают эту выборку для проведения всех необходимых тестов. Но эта выборка не подойдет другим отделам. Почему? Давайте разберемся.

Вам нужны все ваши данные!

По мере того как для решения различных задач используются разные выборки, вы в конечном счете задействуете 100% исходных данных. Не совершайте ошибку, отбрасывая данные, не требующиеся для решения конкретной задачи! Использование выборок не снимает необходимости в сборе и хранении всех релевантных данных. Среда корпоративных данных создается не на основе выборки. Это выборки создаются на основе среды данных.

Возьмем для примера телекоммуникационную компанию. Выборка, включающая 10% клиентов, отлично подходит команде по управлению взаимоотношениями с клиентами (customer relationship management, CRM). Однако вскоре у команды по управлению розничной торговлей возникает необходимость проанализировать эффективность работы розничных точек продаж. Этой команде требуются 10%-ная выборка розничных точек и данные обо всех совершенных в них транзакциях. Эта выборка будет создана совершенно другим способом. В данном случае используется информация не о каждом клиенте, а о каждом магазине. Менеджеру по продукту также может потребоваться 10%-ная выборка данных, относящихся к конкретному товару. Эта выборка не обязательно должна содержать все транзакции, связанные с конкретным клиентом или магазином. Этим трем отделам нужны разные типы выборок.

Дело в том, что любая проблема может потребовать использования всего лишь 10% выборки. Однако, как показано на рис. 7.1, каждая проблема требует выборки, отличной от всех остальных. Со временем, по мере создания разных выборок для решения разных задач, могут быть задействованы все 100% исходных данных. Вот почему необходимо хранить и обеспечивать доступ ко всем данным, несмотря на то что одновременно из них используется не более 10%!



Отдельно взятая проблема может потребовать лишь небольшой выборки данных. В то же время при создании многочисленных выборок для решения разных проблем так или иначе используются все данные

Рис. 7.1. Различные выборки требуют различных данных

Предположения против подсчета статистики

Данная тема — ключевая для понимания разницы между анализом и отчетностью, а также между хорошим и плохим анализом. Представьте себе, что анализ помог обнаружить данные, имеющие статистическую значимость. Аналитик подтвердил, что эти данные актуальны и важны для бизнеса. Теперь он должен сделать предложения по результатам анализа: хороший анализ дает возможность наметить дальнейшие действия. Кроме того, если анализ не поддерживает конкретные действия, то это тоже необходимо документировать.

Хороший анализ упрощает, насколько это возможно, процесс принятия решений для того, кто их принимает, — именно за ним остается последнее слово. Хороший анализ подразумевает выдвижение предположений, а не только сбор статистических данных. Так же как отчет не является анализом, так и простое предоставление статистических данных или другой технической информации им не является.

Недостаточно просто указать, что вариант № 1 превосходит вариант № 2 на 10%. Учитывая все результаты, какое решение следует принять? Хороший анализ будет включать рекомендации. Если вариант № 2 на 10% лучше, чем вариант № 1, то порекомендуйте использовать вариант № 2. В данном простом примере все довольно очевидно.

Однако часто дела обстоят гораздо сложнее. В этих случаях очень полезно предложить возможные дальнейшие действия. Человек, принимающий решение, не должен самостоятельно придумывать, что делать дальше. Ему необходимо предоставить варианты, которые он может принять или отклонить.

Вам нужны профессионалы в области аналитики, а не репортеры!

Работа профессиональных аналитиков заключается в предоставлении результатов анализа и рекомендаций, а не отчетов, данных и статистики. Отчеты обладают ценностью; тот, кто способен анализировать данные и выдавать результаты, необходимые для решения проблемы, — тоже. Большая ценность, однако, обеспечивается благодаря интерпретации этих результатов и разработке плана действий. Именно это превращает отчеты в анализ, а репортеров — в профессиональных аналитиков.

Обзор главы

Самые важные уроки этой главы.

- ▶ Отчет — это не анализ. Формирование отчета — лишь отправная точка для проведения анализа. При правильном использовании анализ и отчетность повышают эффективность друг друга.
- ▶ Анализ подразумевает принятие всех возможных мер для того, чтобы найти основанное на фактах решение бизнес-задачи. Все, начиная от отчетов и заканчивая прогнозными моделями, может играть определенную роль в аналитическом процессе.
- ▶ Хороший анализ отвечает следующим критериям: актуальность, релевантность, объяснимость, полезность и своевременность.
- ▶ Углубленная аналитика выходит за рамки простых вопросов о том, что произошло, когда это произошло и каковы последствия произошедшего. Она также пытается объяснить, почему это произошло и что можно сделать по этому поводу.
- ▶ Один из худших способов использования аналитики в организации — «снятие сливок», то есть отбор подходящих результатов и игнорирование неподходящих. Такое поведение сводит на нет цель и значение анализа.

- Наиболее важная часть любого анализа совершается еще до его начала. Формулировка проблемы может определить успех или неудачу анализа.
- Статистическая значимость — это не то же самое, что важность для бизнеса! Не полагайтесь исключительно на статистические показатели при определении важности результатов анализа.
- Проверка статистической значимости предоставляет лишь вероятность оказаться правым. Свяжите тестируемый уровень значимости с последствиями тех редких ситуаций, в которых будут приняты неправильные решения.
- Даже если у вас есть возможность работать со всей популяцией, это может быть связано с дополнительными расходами и усилиями без каких-либо практических преимуществ. Использование выборки — приемлемая стратегия во многих случаях, в том числе при работе с большими данными.
- Хороший анализ предполагает выдвижение предположений и предложение возможных действий, а не простое предоставление статистических данных и фактов.

Что такое хороший профессионал в области аналитики?

Прежде чем перейти к следующей теме, попробуйте ответить на несколько вопросов. Не волнуйтесь, это просто. Откиньтесь на спинку кресла и подумайте в течение нескольких минут о том, каковы наиболее важные черты аналитика мирового класса. Далее мы будем называть такого человека профессионалом в области аналитики. Он должен быть способен сделать хороший анализ, о чем речь шла в главе 7, и укрощать большие данные. Это высококвалифицированный, подготовленный специалист, который умеет строить прогнозные модели, делать предположения и т. п., а не просто создавать сложные таблицы или отчеты. Запишите от трех до пяти черт, которые считаете наиболее важными для такого специалиста. Ваш список может содержать все, что вы сочтете нужным. После этого продолжайте чтение.

Большая часть читателей этой книги, вероятно, выбрали ответ, по крайней мере частично неверный. Дело в том, что многие общие представления о том, какие черты важны для профессионала-аналитика, неполные или даже неправильные! В этой главе мы узнаем, почему это так, и поговорим о том, что отличает хороших аналитиков от всех остальных. Во-первых, определим, что мы подразумеваем под понятием «профессионал в области аналитики».

Кто такой профессионал в области аналитики?

Людей, которые попадают в категорию профессионалов в области аналитики, именуют по-разному: «аналитик», «специалист в области интеллектуального анализа данных», «специалист по созданию прогнозных моделей» и «специалист по статистике». В последнее время стал довольно популярным термин «ученый в области науки о данных» (data scientist), особенно среди тех, кто много времени тратит на анализ больших данных и работу с такими инструментами, как MapReduce. Мы будем называть всех этих людей «профессионалами в области аналитики».

Дело в том, что основные навыки людей, о которых идет речь, имеют больше сходств, чем различий. День за днем каждый из этих специалистов использует данные для решения бизнес-задач. Инструменты или алгоритмы могут быть различными, однако профессионал в области аналитики, работающий в одной области, может при необходимости подстроиться под работу в другой области. Как мы увидим далее, хорошего профессионала в области аналитики от всех остальных отличают не инструменты, алгоритмы или данные, которыми он обычно пользуется.

Наиболее показателен тот факт, что новое сообщество ученых в области науки о данных не сильно отличается от сообщества традиционных аналитиков. И те и другие пытаются найти новые эффективные способы использования данных для решения бизнес-задач. Тот факт, что ученые в области науки о данных, как правило, используют другие наборы инструментов, языки программирования и наборы данных, не меняет их основных целей и задач. В игру вступают одни и те же базовые навыки и компетенции.

Для того чтобы хороший аналитик стал хорошим ученым в области науки о данных и наоборот, требуется только дополнительное обучение. Любой хороший профессионал в области аналитики легко может научиться использовать новый язык программирования или инструмент. И любой хороший специалист непременно ухватится за возможность узнать о новом источнике данных и способах его применения.

Те, кто считает себя профессионалом в области аналитики независимо от того, как он себя называет, должны согласиться со всем, что сказано в этой главе. Те, кто взаимодействует с этими специалистами, тоже с этим должны согласиться. Очень важно, чтобы профессионалы в области аналитики понимали, как много у них общего.

Распространенные заблуждения о профессионалах в области аналитики

Большая часть людей, которых попросят составить список важных черт для профессионала в области аналитики, назовут наличие какой-либо ученой степени. Считается, что настоящий специалист нуждается в дипломе по статистике, математике, информатике, исследованию операций и т. п. Нередко предполагается, что успех требует магистерской или докторской степени. Как правило, большое значение придают наличию опыта программирования: говорят, что хороший специалист должен уметь использовать один из множества языков программирования, предназначенных для анализа. Логика данного критерия заключается в том, что такие люди владеют определенными инструментами и хороший аналитик должен уметь эффективно их применять.

Однако это не так. Дело в том, что большинству профессионалов в области аналитики необходимы глубокие знания математики и статистики. Наличие диплома не обязательно. Все эти знания можно приобрести в процессе работы или другим образом, отличным от формального обучения. Хорошему специалисту, конечно, нужны навыки программирования, поскольку использование всех основных аналитических инструментов требует определенного уровня знаний в области программирования. Тем не менее навыки программирования еще не гарантируют успех.

Эти ответы неправильны по той причине, которую можно сформулировать с помощью фразы, используемой в математических доказательствах: «Необходимое, но недостаточное условие». Наличие навыков в области статистики, математики и программирования, безусловно, необходимо, но их недостаточно, чтобы человек стал профессионалом в области аналитики. Для этого требуется гораздо больше. Наличие базовых знаний по математике и программированию — данность. Эти факторы важны, но не они отличают хорошего специалиста от всех остальных. Это лишь отправная точка.

Если менеджер по найму слишком много внимания уделяет техническим навыкам и академическим знаниям, то рискует принять на работу человека, который сконцентрирован именно на них, а не на общей картине, связанной с проведением содержательного анализа. Важно, чтобы при поиске профессионалов в области аналитики организации опирались и на другие критерии. В конце концов, никому

не нужен «помешанный на статистике» сотрудник, сидящий в углу и круглосуточно запускающий сложные алгоритмы. Это не приводит к успеху.

Компании нужен специалист, который будет частью команды. Он должен понимать бизнес-задачи, которые стоят перед организацией, и уметь эффективно производить анализ, необходимый для их решения. Вы не сможете укротить большие данные без первоклассного талантливой специалиста. Теперь обсудим критерии, которые необходимо учитывать при поиске такого специалиста.

Каждый хороший профессионал в области аналитики — исключение

Меня долгие годы удивляло, что практически каждый хороший профессионал в области аналитики, с которым я был знаком, представлял собой в некотором роде исключение из правил. Я знаю многих людей, которые подтверждают мои наблюдения. Что я имею в виду? Если посмотреть на список качеств, обычно считающихся главными признаками хороших специалистов, то самые лучшие из них будут нарушать один или более из приведенных в этом списке пунктов! Для начала посмотрим, почему некоторые черты не столь уж важны, как может показаться, прежде чем перейдем к более значимым чертам.

Образование

Много лет назад я работал с одним из лучших профессионалов в области аналитики, которых когда-либо знал. Назовем его Барт. Он присоединился к компании раньше меня, поэтому я не знал о его образовании, когда начал там работать. Однако довольно быстро я понял, что этот парень действительно знал свое дело. Я мог подойти к нему и получить ответ на вопрос, связанный с программированием, со статистикой и с делами компании, в которых пытался разобраться, будучи новым сотрудником. Самое главное, он мог помочь мне понять бизнес наших клиентов.

Со временем я узнал, что у него была «всего лишь» степень бакалавра по бизнесу. У Барта не было других ученых степеней. Не было дипломов по статистике или математике. Он окончил курсы по стати-

стике, предусмотренные его программой бакалавриата. Барт не имел формального образования в области программирования. Он научился программировать самостоятельно.

Когда Барта приняли на работу, он окончил несколько курсов и получил необходимые ему основные знания по статистике от своих более опытных коллег. Он прочитал несколько книг. Барт научился программировать в процессе работы. Он сделал из себя одного из лучших профессионалов в области аналитики, которых я когда-либо знал. Однако у него не было формального образования и технической подготовки, наличия которых можно было бы ожидать. Это не имело значения. Но именно этому человеку вы захотели бы поручить укрощение больших данных. Не заикливайтесь исключительно на формальном образовании. Лучше постарайтесь выяснить, изучил ли данный специалист все, что требуется для выполнения своей работы.

Отраслевой опыт

Очень часто компании и менеджеры по найму озабочены тем, в какой отрасли аналитик или кто-либо другой имеет опыт работы. Если человек работал в сфере телекоммуникаций, то многие полагают, что он не может принести дополнительную пользу в банковской сфере. Если это специалист в банковском деле, считают, что от него будет мало пользы в сфере производства. Если аналитик работает в сфере производства, они даже не подумают о том, чтобы взять его на работу в сфере розничной торговли.

Это несправедливо. Если необходимо сделать выбор между двумя одинаково квалифицированными кандидатами, один из которых знает отрасль, а второй — нет, конечно, имеет смысл принять на работу того, кто имеет опыт работы в данной отрасли.

Однако выбор редко бывает таким простым. Допустим, вам нужно сделать выбор между двумя кандидатами. Один из них — посредственный аналитик, который очень хорошо знает отрасль. Другой — выдающийся профессионал в области аналитики, который имеет опыт в другой отрасли и ничего не знает о вашей. Смело выбирайте второго. Хороший профессионал в области аналитики в одной отрасли может изучить другую и стать профессионалом и в ней. Посредственный аналитик, вероятно, таким и останется. Кроме того, новый

взгляд со стороны может быть весьма полезным. В каждой отрасли есть определенные способы ведения дел. Команда может многому научиться у того, кто пришел из другой отрасли.

Ищите за пределами отрасли

Время от времени, нанимая профессиональных аналитиков, ищите людей за пределами вашей отрасли. Хороший специалист быстро вникнет в ваш бизнес. Он обогатит вашу команду новыми идеями и подходами. Сотрудник конкурирующей компании, который уже умеет делать именно то, что вам нужно, — отличный выбор, однако время от времени следует его разнообразить.

В связи с этим мне вспоминается история Марка. В течение нескольких десятилетий он работал главным образом в банковской индустрии. Наша команда испытывала дефицит ресурсов, и нам нужна была помощь в работе над проектом в сфере розничной торговли. Все соглашались с тем, что Марк был очень, очень хорошим профессионалом в области аналитики. Но мог ли он справиться с проектом, связанным с розничной сферой, учитывая его опыт в банковской отрасли?

Хорошие профессионалы в области аналитики согласятся с тем, что, если кто-то смог разобраться в одной отрасли, он сможет сделать это и в другой. Вероятно, ему придется несколько изменить свой образ мыслей, познакомиться с новой терминологией и научиться вычислять некоторые новые метрики, однако он может добиться успеха. Это справедливо и в отношении многих других дисциплин. У Марка появилась возможность поработать над розничным проектом. Он согласился потратить время, чтобы познакомиться с отраслью, а также тесно сотрудничать с экспертами в области розничной торговли. Даже этот первый проект оказался успешным, и через несколько месяцев Марк уже встречался с другими розничными клиентами, которые были уверены, что он занимался розничной торговлей много лет! Это произошло благодаря тому, что Марк вник в основные концепции бизнеса и сумел применить знания, полученные в процессе работы в банковской отрасли, в сфере розничной торговли. Он был мотивированным, творческим и умным человеком. Вот это действительно имело значение.

Остерегайтесь использовать «список»

Несколько лет назад сотрудник кадровой службы компании, в которой я работал, сказал мне: «У нас появились новые правила, так что нужно

обновить список необходимых и предпочтительных черт соискателей. С этого момента все, что вы укажете в качестве требования, будет абсолютным». Другими словами, если в описании вакансии указано, что кандидат должен иметь степень бакалавра или выше в области статистики, то нельзя проводить собеседование или нанимать того, у кого ее нет.

Я подумал об этом и отправил в кадровую службу пересмотренный длинный список предпочтительных характеристик. Мой список необходимых требований включал только один пункт: степень бакалавра в любой области. Единственная причина, по которой я указал степень бакалавра, состояла в том, что у кандидата должен быть определенный уровень образования. По правде говоря, этот список мог показаться слишком суровым, даже учитывая наличие одного требования.

Сотрудник кадровой службы позвонил и сказал: «Эй, Билл, ты не ошибся? Ты, по сути, не указал ни одного требования. Неужели нет определенных характеристик, которые тебе абсолютно необходимы?» Я сказал им то же, о чем шла речь выше: «Честно говоря, каждый человек в моей команде не соответствует по крайней мере одному или двум критериям, которые обычно считаются необходимыми. Если я укажу какое-то требование и не смогу сделать исключение, то мне нельзя помещать это в список. Я не могу рисковать потерей хорошего человека из-за того, что описание вакансии высечено в камне. Я лучше буду использовать неоднозначное описание вакансии, чтобы иметь возможность найти нужного человека».

Нанимайте знания и навыки, а не галочки

Принимая на работу профессионала в области аналитики, начните с характеристик, касающихся опыта работы и образования. Наличие некоторого количества галочек — первый этап. Но этого недостаточно. Профессионализм в области аналитики напрямую не связан с техническим образованием. Хорошие специалисты отличаются от остальных в меньшей степени техническими навыками, чем другими характеристиками, о которых пойдет речь далее.

Что еще нужно ценить в профессионалах-аналитиках

Поговорим о качествах, которые отличают хороших аналитиков. Впрочем, они важны и в других профессиональных областях. Для поддержания налаженных аналитических процессов требования могут быть

слегка снижены, однако привлечение и удержание талантливых сотрудников имеет первостепенное значение, когда организация стремится выйти на новый уровень и создать новые аналитические процессы. Еще выше планку необходимо поднять при необходимости укротить большие данные и создать новые, инновационные, меняющие бизнес аналитические процессы.

Ответственность

Способность взять на себя обязательства и выполнить их — преимущество в любой профессии. Есть люди, которые сделают все возможное для успешной и своевременной сдачи проекта. Но есть и такие, кто готов вложить в дело только определенное количество усилий. В вашей организации вы, вероятно, знаете людей, на которых можно и нельзя положиться. Любой хороший профессионал в области аналитики будет брать на себя обязательства и выполнять их. К счастью, данная черта, как правило, выявляется в ходе интервью, когда кандидат рассказывает о его прошлой работе и успехах. Слушайте внимательно, чтобы определить уровень его ответственности.

Мало что можно добавить об ответственности. Эта характеристика понятна и важна в любой области, в том числе в сфере аналитики.

Творческий подход

Большинство людей вряд ли связывают творчество с профессиональной аналитикой. Считается, что аналитики используют определенные заранее статистические формулы. Аналитик работает по книге, и в этом нет никакого творчества. Так ли это?

Не совсем. Суть в том, что каждая бизнес-задача уникальна и имеющиеся данные для решения каждой проблемы часто сложны и неполны. Профессионалу в области аналитики необходимо определить, как можно использовать эти данные для решения конкретной задачи. Это требует творческого подхода. Не существует книги или набора правил, который позволял бы принимать правильные решения.

Каждый раз, проводя анализ, специалист сталкивается с непредвиденными трудностями. Иногда они незначительны, а порой существенны. Бывает, перед аналитиками возникают очень сложные задачи.

Для того чтобы найти новые решения, и требуется творчество. Проблема может быть связана с данными или с одним из аспектов бизнеса, который не был до конца понят раньше. Творческий подход помогает преодолеть эти барьеры и получить нужный конечный результат.

Да, творчество — не самая распространенная среди аналитиков черта, однако не стоит недооценивать ее значение. Творчество — важный фильтр. Поговорите с десятком человек, и вам повезет, если хоть некоторые из них проявят признаки творчества. Некоторые организации для оценки творческого потенциала используют личностные тесты или предлагают кандидатам решить, казалось бы, случайно выбранные задачи. Я часто оцениваю творческий потенциал, задавая вопрос о том, как кандидатам удавалось справляться с возникавшими сложностями. Творческий специалист сможет рассказать вам хорошую историю. Аналитик, не обладающий творческим потенциалом, просто перечислит список предпринятых шагов, с помощью которых он пытался решить эту проблему.

Чистые данные существуют только в учебниках

Уместно ли обсуждение темы чистых данных в разделе, посвященном значению творческого подхода для хорошего профессионального аналитика? Да, поскольку аналитикам часто приходится проявлять творчество при обработке данных. Дело в том, что данные никогда не бывают такими чистыми, какими должны быть. В них всегда есть пробелы, несоответствия и просто ошибки. Они также почти всегда противоречат некоторым предположениям, сделанным в процессе разработки плана анализа.

В ходе обучения используются точные, чистые и полные данные. Если есть не укладывающиеся в общую картину точки данных, можно определить причину этого и произвести необходимую коррекцию. Все учащиеся наивно полагают, что примеры в учебнике отражают то, с чем им предстоит столкнуться в мире бизнеса. В бизнесе все происходит иначе. Данные никогда не бывают столь же замечательными, как в школе. Обозначения пола могут содержать что-то вроде H в дополнение к ожидаемым M, F и U (unknown, если пол не указан). У клиента продуктового магазина может быть указана покупка на сумму \$10 000 000. Может быть продан продукт с несуществующим товарным кодом.

Такие ситуации порождают серьезные проблемы: что делать, если вы видите несоответствие в данных? Следует ли проигнорировать клиента с нереальным объемом покупки? Нужно ли заменить обозначение H на U? Можно ли определить правильный код продукта? Определение наиболее эффективного способа использования доступных данных — один из самых трудных этапов любого аналитического процесса, который может потребовать применения творческого подхода. Когда специалист понимает, что данные недостаточно полны, чтобы найти ответ на поставленный вопрос, приходится проявить творческий подход, чтобы найти наиболее эффективный способ использования доступной информации. Например, игнорировать или скорректировать некоторые данные. Бывает необходимо добиться некоторых немедленных результатов, а затем сосредоточиться на дальнейшем улучшении.

Стремитесь к улучшению, а не к совершенству

При решении бизнес-задач следует ставить цель улучшить результаты, а не достичь совершенства. Это имеет решающее значение. Легко заикнуться на дополнительной работе по очистке данных. Хорошие профессионалы в области аналитики фокусируются на улучшении результатов и на том, чтобы выжать максимум из имеющихся данных. Результаты могут не быть совершенными, однако если они обеспечивают более эффективное принятие решения, то это не страшно.

Например, при анализе данных карт лояльности приходится мириться с тем, что данные никогда не бывают совершенными. Даже самые лучшие клиенты иногда забывают использовать свои карты. Это означает, что «полная» история расходов каждого клиента на самом деле не может считаться полной. Однако не все потеряно! Действительно хорошие клиенты используют свои карты большую часть времени. При этом данных будет достаточно для того, чтобы понять этих клиентов. Отсутствие некоторых данных не мешает проведению анализа. Конечно, на основе частичной информации могут быть сделаны не совсем верные выводы, однако данных будет достаточно для принятия правильных решений. Хорошие специалисты это понимают.

Достаточно чистые данные

Независимо от качества данных хорошие аналитики пытаются определить, достаточно ли они чистые. Можно ли с их помощью получать достоверные результаты? Можно ли, опираясь на них, ожидать полу-

чения реальной выгоды? Если ответ положительный, то такие данные используют. Данные могут не быть совершенными, но должны быть достаточно хорошими, чтобы позволить принять решение. Настоящие профессионалы в области аналитики умеют применять творческий подход при проверке данных на предмет достаточной чистоты.

Рассмотрим источник данных, который широко используется, но полон ошибок. Демографические данные были доступны в течение десятилетий. В среднем поставщикам демографических данных удается получить довольно точную информацию о домохозяйствах. Тем не менее данные, относящиеся к конкретному домохозяйству, могут быть неправильными, поскольку в процессе компиляции данных выдвигаются некоторые предположения. Это не делает данные бесполезными. Закономерности и тенденции высокого уровня можно считать верными, даже если данные о конкретных домохозяйствах в рамках этих тенденций неидеальны. Маркетологи с успехом использовали такие данные, несмотря на их «несовершенство». Существуют творческие способы корректировки известных предубеждений и проблем. Если данные игнорируются из-за имеющихся в них ошибок, то их ценность будет востребована не в полной мере.

В корпоративной среде профессиональный аналитик сможет, проявив творческий подход, использовать источники корпоративных данных для повышения ценности. Имеющиеся данные следует рассматривать как стакан с ценностью, наполовину полный, а не наполовину пустой. Этот навык еще более ценен при работе с большими данными, учитывая факты, которые приведены в первой главе. Большие данные часто недостаточно чистые и содержат постороннюю информацию, которую необходимо отфильтровать.

Деловая смекалка

Хорошие профессионалы в области аналитики должны разбираться в бизнес-модели, в рамках которой они работают, а также в том, как аналитика может помочь решить соответствующие бизнес-задачи. Хорошие аналитики умеют сосредоточиться на том, какие показатели и результаты важны с точки зрения бизнеса, а какие — с технической точки зрения. Они готовы потратить время и приложить усилия для достижения этой степени понимания. Независимо от уровня врожденной деловой смекалки человеку требуется проявить интерес и внимание,

чтобы развить ее. Если у человека нет интереса и желания разобраться в вашем бизнесе, то он не сможет стать для вас хорошим специалистом.

Обратите внимание на то, что деловая смекалка и отраслевой опыт — не одно и то же. Отраслевой опыт — это накопленные знания. Деловая смекалка предполагает наличие иных навыков. Человек с хорошей деловой смекалкой может с легкостью разобраться в любой новой отрасли. Хорошие специалисты вроде Марка, о котором я рассказывал, могут применить смекалку к разнообразным ситуациям и задачам. Проводя собеседование, старайтесь выяснить у кандидатов, почему они приняли то или иное решение в процессе работы над предыдущими проектами. Если у них хорошо развита деловая смекалка, то наряду с техническими доводами они приведут и практические соображения, относящиеся к бизнесу. Кроме того, они прокомментируют, почему при решении конкретной бизнес-задачи определенные проблемы были в большей или меньшей степени актуальны. Профессиональный аналитик, не обладающий деловой смекалкой, сосредоточится в первую очередь на технических аспектах.

Странный гибрид

Выдающиеся профессионалы в области аналитики представляют собой странный гибрид. То они должны сосредоточиться на технической стороне вопроса, как настоящие ИТ-специалисты, то им приходится проявить деловую смекалку, какой обладают настоящие бизнесмены. Эти черты сочетать довольно трудно. Вот почему трудно найти действительно выдающихся аналитиков.

Подходящий уровень детализации

Один из аспектов деловой смекалки — умение связать результаты анализа с уровнем детализации принимаемого решения. Что это значит? Допустим, специалист в области бизнеса поручает аналитику улучшить маркетинговую кампанию. Если он сумеет создать такую модель, эффективность которой на 2% превысит эффективность используемой в настоящее время, то ей будет отдано предпочтение. В этом случае перед аналитиком была поставлена определенная цель: удостовериться, что текущие результаты можно улучшить по крайней мере на 2%.

Представляя результаты, заявит ли он, что уровень эффективности модели превысил базовый уровень на 5,32526%? Надеюсь, что нет. Это особенно важно, если погрешность составляет плюс-минус 2%. Кого волнует, что точечная оценка равна 5,32526%, когда погрешность

составляет плюс-минус 2%? Указание тысячных долей процента только отвлекает внимание. Самое главное — донести, что результаты могут быть улучшены на 5% плюс-минус 2%. В худшем случае результат может быть улучшен чуть больше, чем на 3%, так что эта модель будет явным победителем, учитывая изначальный ориентир в 2%. Это все, что нужно знать бизнесменам. Хороший аналитик не будет загружать бизнес-группу подробностями сверх того, что им необходимо и ценно. Деловая смекалка поможет ему определить, что и как следует преподнести.

Другой пример связан с прогнозированием спроса. Несколько лет назад один из поставщиков доказал, что его прогнозы были гораздо более точными, чем у конкурентов. Он показал, что в среднем организации необходимо иметь в наличии только три дополнительные единицы товара, а не четыре, как рекомендовали его конкуренты. Спонсоры проекта были довольны, но задали вопрос, который несколько остудил энтузиазм. Они отметили, что объем заказа должен быть кратным шести единицам! Учитывая этот факт, как они могли эффективно использовать любой из прогнозов? Когда уровень детализации действия составляет шесть единиц, все, что указывает на большую степень детализации действия, бесполезно. Специалист, обладающий хорошей деловой смекалкой и умеющий правильно формулировать проблему, сможет заранее определять такие ограничения и вносить соответствующие корректировки.

Концентрация на самом важном

Данные часто противоречат предположениям. Например, при создании многих моделей требуется сделать предположение относительно нормального распределения. Теоретически, когда такие предположения не оправдываются, возникают большие проблемы. Однако на практике при наличии тесной взаимосвязи между двумя факторами она проявится вне зависимости от используемого метода. Означает ли это, что оценки параметров и прогнозируемый эффект будут идентичными при различных вариантах моделирования, если предположения окажутся неверными? Нет, но это значит, что важные факторы, как правило, будут считаться важными при использовании любых методов, даже если предположения не оправдаются. Если в высоком уровне детализации нет необходимости, то можно ограничиться приблизительными значениями.

Бывает ли так, что совершенная U-образная кривая настолько противоречит предположению о наличии линейной зависимости, что линейная регрессия показывает полное отсутствие какой-либо связи между двумя переменными? Да, бывает. Однако вероятность этого невелика и в большинстве случаев наличие взаимосвязи будет обнаружено. Если тому, для кого делается анализ, нужны приблизительные данные для принятия решения типа да/нет, то данные и модели должны быть достаточно точными, чтобы однозначно ответить на поставленный вопрос. Хороший аналитик знает, когда следует повысить или понизить уровень точности, исходя из существующих требований. Данные, которые полностью противоречат предположению о наличии линейной зависимости, представлены на рис. 8.1. Тем не менее линия регрессии эффективно отражает суть существующей взаимосвязи, если все, что требуется, — это установить факт наличия зависимости двух факторов друг от друга.

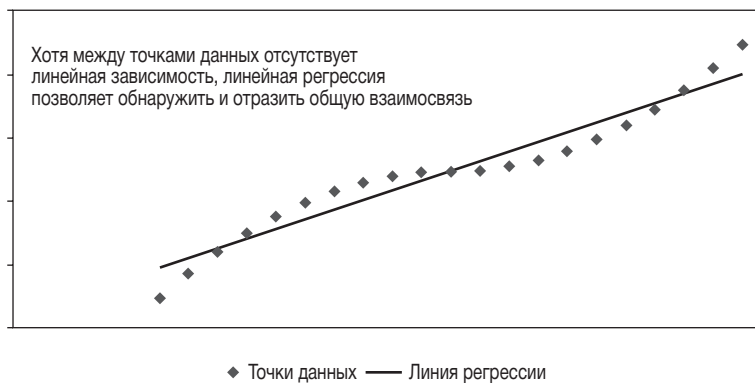


Рис. 8.1. Линейное представление нелинейной зависимости

Культурная осведомленность

Во многих отраслях существует тенденция использования аутсорсинга, особенно в развивающихся странах. Хорошо это или плохо, но это касается и сферы аналитики. Мы не будем рассматривать политические и философские аргументы о том, хорош ли аутсорсинг или плох с экономической или моральной точки зрения. Оставим это для другого случая. Сейчас важно разобраться в том, может ли сегодня аутсорсинг удовлетворить потребности бизнес-аналитики.

Большая часть офшорных провайдеров на момент написания этой книги фокусируются на технических навыках и на предоставлении

услуг технически подготовленной команды. Они подчеркнут, что у них есть 25 докторов наук в области статистики, которые умеют использовать все существующие программные пакеты. Просто поставьте перед ними задачу, и они ее решат. Как мы уже отмечали, проблема заключается в том, что технические навыки — лишь необходимый минимум, которым должен обладать хороший специалист. Кроме того, очень трудно проявить деловую смекалку, если вы ни разу не видели в действии тот бизнес, который анализируете.

Услуги офшорных провайдеров могут пригодиться для решения хорошо определенных аналитических задач. Однако с использованием офшорных ресурсов для обеспечения полного комплекса аналитических услуг связаны огромные недостатки. Рассмотрим типичную ситуацию, когда сотрудник находится на другом конце света, вас разделяют несколько часовых поясов и языковой барьер. Эта ситуация создает проблемы сама по себе. Теперь задумайтесь об огромных культурных различиях и отсутствии у офшорных провайдеров представления о том, как обстоят дела в стране, с которой они работают.

Обратите внимание, что эти трудности существуют независимо от того, кто осуществляет удаленную поддержку. У американцев будет столько же проблем при предоставлении аналитических услуг организации в Индии, которую они никогда не видели, как и у индийских аналитиков, работающих с американской организацией, которую ни разу не видели они.

Коллега рассказал мне показательную историю об организации, работающей в сфере бакалейной торговли, которая поручила офшорному провайдеру провести анализ своей категории кормов для домашних животных. Представьте себе банки и пакеты с собачьим кормом, на которых изображена счастливая собака. Когда организация получила результаты анализа, из формулировки отчетов, а также из устного доклада стало понятно, что аналитики абсолютно неверно поняли, что собой представляла анализируемая категория товаров. Результаты касались не корма для домашних животных, а консервов из мяса собаки! Вы уже поняли, к чему я веду? Офшорная команда посчитала, что собака, изображенная на этикетке, довольна не потому, что ей предстоит полакомиться содержащимся в банке кормом, а потому, что ее саму запихали в эту банку, чтобы мы смогли съесть ее на ужин!

Такая же проблема легко могла возникнуть, если бы организации поменялись ролями. Очень сложно поддерживать высокий уровень

деловой смекалки, когда вы совершенно незнакомы с бизнесом и культурой, в которой работает данная организация. Могут ли офшорные ресурсы быть полезными? Да, если они используются надлежащим образом. Не перебрасывайте аналитическую бизнес-задачу через стену в надежде на то, что офшорная команда с чисто технической подготовкой сможет разработать аналитическую стратегию, интерпретировать результаты и донести их до вас. Чтобы добиться успеха, вам необходимы действительно хорошие, обладающие деловой смекалкой профессионалы в области аналитики, которые будут руководить процессом непосредственно из штаб-квартиры.

Навыки презентации и коммуникации

Навыки презентации и коммуникации имеют решающее значение для многих профессионалов, включая аналитиков. Независимо от того, насколько хорошо специалист умеет находить правильные результаты, при продвижении по карьерной лестнице ему потребуется не только производить анализ, но и уметь связать результаты в единую убедительную и краткую историю. Хороший аналитик способен заинтересовать далеких от технических вопросов людей и донести до них информацию в понятных им терминах. Он придумает интересную историю, а не просто перечислит ряд статистических показателей и фактов.

Профессиональный аналитик не может войти в комнату и сообщить бизнес-аудитории о диагностике коллинеарности, подробной статистике модели и других, сугубо технических деталях. Он должен войти и сказать: «Вот что мы нашли, вот почему это важно и вот что вы можете сделать». Он должен перечислить преимущества, которые могут быть получены в случае принятия рекомендуемых мер. Увеличится ли объем продаж? Повысится ли прибыль? В конце дня бизнесмены хотят знать о том, как анализ может им помочь, а не обо всех технических моментах, связанных с тем, как были получены представленные результаты.

Результаты должны быть донесены кратко и четко. Чрезвычайно важны навыки письменного общения вне зависимости от того, создает ли специалист слайд-шоу или письменный документ, навыки речевого общения и презентационные навыки, будь то создание официальной презентации или проведение неформального обсуждения в офисе.

Не каждому специалисту приходится предстать перед аудиторией на конференции или на заседании исполнительного комитета.

По крайней мере не на ранних этапах карьеры. Тем не менее каждому профессионалу потребуется умение выступать перед спонсорами проекта и/или его собственным начальством в офисе или конференц-зале и доносить полученные результаты. Оценить презентационные навыки кандидатов можно, предложив им провести презентацию в процессе собеседования. Можно задать им общую тему или позволить выбрать что-то самостоятельно. Вы увидите их в действии и под давлением и сможете оценить их навыки общения за несколько минут.

Устройте тест-драйв

Для того чтобы оценить презентационные навыки аналитика, можно попросить его провести презентацию в процессе собеседования. Вы увидите, что собой представляет данный кандидат, и сможете сказать, что ищете того, кто может стать отличным аналитиком в вашей организации.

Результаты — не самый важный фактор успеха

Это утверждение может вас удивить. Однако самый важный фактор при определении вероятности успеха данного аналитического проекта — не качество результатов анализа. В идеальном мире было бы именно так. А в реальном мире, в котором мы живем, дело обстоит иначе. Во-первых, давайте условимся, что получение правильных результатов обязательно. Это принципиально важно, и любой аналитик должен гарантировать точность результатов. Тем не менее в конце рабочего дня с точки зрения людей, спонсирующих аналитические проекты, на сами результаты приходится не более половины критериев, определяющих, сочтут ли они данный проект успешным. Так что еще может иметь значение?

По крайней мере 50% успеха проекта зависит от того, насколько хорошо аналитик сделает презентацию и документирует результаты. Может ли он эффективно позиционировать свои результаты? Умеет ли интерпретировать показатели так, чтобы донести смысл до своей аудитории и люди могли принять соответствующие меры? Важность этого сложно переоценить. Хороший специалист не может сосредоточиться только на аналитике, как бы ему этого ни хотелось. Он также должен оставить время на объяснение, позиционирование и продажу результатов бизнесменам, по просьбе которых проводился анализ.

Бизнес-команде нет дела до десятидневной работы и всех трудностей, которые пришлось преодолеть. Они хотят знать только

о результатах. Аналитик должен эффективно донести информацию об итогах проведенного анализа, в противном случае эти результаты практически бесполезны. Опять же, получение правильных результатов — необходимое, но недостаточное условие, чтобы проект считался успешным. Хороший профессионал понимает это и уделяет представлению должное внимание.

Дело в представлении, глупыш!

Для того чтобы научиться представлять результаты в удобоваримой форме, требуются время и старание. Иногда аналитикам будет казаться, что в их представлении слишком много «воды». Хотя детали, подтверждающие результаты, должны быть доступны, к ним стоит обращаться только по необходимости. Если обсуждение станет слишком детальным, то глаза слушателей потускнеют, они отключатся и не будут использовать результаты. Хороший специалист донесет результаты так, чтобы заинтересовать спонсоров.

Урок из области рекламы

Профессиональные аналитики любят производить измерения. Они стараются делать все для того, чтобы можно было доказать, приводят усилия к улучшению результатов или нет. В сфере прямого маркетинга это проявляется особенно ярко. Например, аналитики использовали модель для создания списка людей, которым необходимо отправить электронные письма, позвонить или связаться каким-то другим образом. В данном случае полученный результат легко измерить. Если он положительный, то можно продолжить делать то же самое; если нет — попробовать что-то другое.

Одна из самых больших расходных статей бюджета множества компаний — затраты на рекламу на телевидении, радио, в газетах и т. д. Такие СМИ играют определенную роль, однако эффект их использования практически невозможно точно измерить. Оценка эффективности рекламы — очень непростое дело. Методологии оценки повышения уровня продаж за счет рекламы на телевидении, радио и в печати в лучшем случае неоднозначны. Эффективность методологий, используемых на более низком уровне, например на уровне отдельного магазина, немногим выше. Тем не менее к рекламе прибегают часто, хотя существуют другие, легко измеримые мероприятия, в пользу которых можно перераспределить бюджеты. Почему так происходит?

Одна из причин заключается в том, что, когда рекламодатель, использующий прямой маркетинг, хочет провести анализ с целью повышения

эффективности таргетированной рекламы, в этом нет ничего особенно захватывающего. Такой анализ выявляет тех, кто с большей степенью вероятности откликнется на предложение, а затем организация нацеливает на этих людей свою рекламу в надежде поднять уровень продаж. Конечно, вопрос о повышении объема продаж заставляет загораться глаза людей, однако в самом сюжете нет ничего особенно интересного.

Что делают рекламные агентства, представляя свои планы? Создают мультимедийную презентацию. Они используют музыку, видео и цепляющие фразы. Они так заинтересовывают аудиторию своими планами, что слушатели готовы подписать договор. Даже тот факт, что точно измерить эффективность невозможно, не имеет значения, поскольку аудитория прониклась видением того, что предлагает сделать данное рекламное агентство.

Смысл не в том, чтобы упрекнуть рекламные агентства (пожалуйста, не пишите гневных писем). Напротив, это скорее комплимент! Дело в том, что эффективность рекламы измерить намного сложнее, чем эффективность других действий, и тем не менее именно на рекламу приходится огромная доля расходов. Это частично объясняется способностью рекламщиков убеждать спонсоров прибегнуть к их услугам. Рекламные агентства в полной мере понимают и используют навыки презентации и общения. Профессиональный аналитик, желающий стать выдающимся, может извлечь определенные уроки, изучив деятельность рекламных агентств. Представьте, каким эффективным может стать проект, если измеримые действия и связанная с ними аналитика будут дополнены воодушевлением, свойственным рекламной деятельностью.

Интуиция

Это, вероятно, наиболее сложная для описания характеристика. Интуицию невозможно оценить, пока вы не увидите человека в действии. Для наших целей мы определим интуицию как возможность аналитика каким-либо образом понимать или чувствовать то, что следует делать дальше. Когда возникает препятствие, аналитик сидит и продумывает сложившуюся ситуацию. Он определяет четыре возможных решения проблемы. Как он делает выбор? Ощущает ли интуитивно, какой из вариантов самый лучший? Достигает ли он успеха при выборе того или иного варианта в большинстве случаев? Или пробует различные

варианты, прежде чем разработать план? Выдающийся специалист способен в большинстве случаев выбрать самый лучший путь.

Есть очень хорошая книга, которую я рекомендую вам прочесть. Она называется «Будущее за правым полушарием» (A Whole New Mind), автор Дэниел Пинк*. В этой книге очень доступно изложены некоторые темы, о которых шла речь.

Во многих отношениях интуиция — врожденное качество. Однако ее можно развить и усовершенствовать. Интуиция представляет собой сочетание умения решать проблемы и опыта ведения дел в подобных ситуациях в прошлом. Основываясь на этих подходах и опыте, необходимо понимать, когда следует применять предыдущий опыт, а когда его следует подстроить под новую ситуацию.

Развитая интуиция — важная черта для хорошего профессионала в области аналитики, однако во время собеседования очень трудно оценить уровень интуиции кандидата. Некоторые из кажущихся уместными критериев могут не сработать, поскольку они слишком субъективны и неоднозначны. Со временем вы сможете оценить интуицию аналитика на основании того, как он работает и справляется с проблемами.

Искусство или наука

Аналитика — не только наука. Это также искусство. Хороший анализ сочетает в себе серьезный научный подход с большой долей артистизма. Артистизм необходим, когда предстоит справиться с необычными проблемами, создать убедительную презентацию и доходчиво интерпретировать результаты. Хороший профессионал в области аналитики должен развиваться как в области науки, так и в области искусства, поскольку он не только ученый, но и художник!

Рассмотрим для примера кластерный анализ — распространенный набор алгоритмов, используемых для создания моделей сегментации. Не существует простых общепринятых метрик, определяющих правильный ответ. Моделирование сегментации относится скорее к искусству. Профессионалы в области аналитики, которым часто приходится создавать модели сегментации, придерживаются в этом собственных принципов. Например, я при оценке таких моделей следую определенному процессу. Я знаю, когда мне нужно сделать это, куда я направляюсь

* Пинк Д. Будущее за правым полушарием. М. : РИПОЛ классик, Открытый Мир, 2009. Прим. ред.

и какие закономерности ищу. Однако мне было бы сложно доходчиво объяснить кому-то другому некоторые стадии этого процесса. Точно так же другим людям сложно объяснить свои методы мне. Каждый человек по-своему оценивает модели сегментации, и здесь присутствует значительная доля искусства.

Доверие — важный фактор в работе профессиональных аналитиков, и оно как никогда важно, когда речь заходит о специалистах, чья работа включает большую художественную составляющую. При отсутствии однозначных метрик, позволяющих сделать тот или иной выбор, бизнес-спонсорам приходится доверять интуиции аналитика и его искусству. Чтобы заслужить безоговорочное доверие, потребуется много времени. Выдающийся профессионал в области аналитики будет завоевывать это доверие и со временем станет настоящим советником для своих бизнес-партнеров.

Профессионалы в области аналитики как художники

Так же как два разных художника могут совершенно по-разному интерпретировать сцену и при этом создать замечательные произведения искусства, так и два разных человека в результате проведения анализа могут придумать различные подходы, по-своему замечательные. В этом заключается артистическая составляющая аналитики. Некоторые алгоритмы оставляют мало возможностей для творчества, однако оно всегда лежит в основе решений о том, как лучше определить проблему, разработать аналитический процесс и обработать имеющиеся данные, чтобы прийти к решению. Хорошие профессионалы в области аналитики в равной степени и художники, и ученые.

В последнее время в сообществе аналитиков часто обсуждается роль ученого в области науки о данных в организациях. Мы уже упоминали, что сегодня нет большой разницы между тем, что делает ученый в области науки о данных, и тем, что всегда делали профессионалы в области углубленной аналитики. Появляются новые инструменты вроде MapReduce, которые должен освоить аналитик, чтобы называться ученым в области науки о данных, однако освоение новых инструментов для таких людей — привычное дело. Имеет ли ученый в области науки о данных какие-либо новые инструменты? Да. Стоит ли перед ним совершенно новая цель? Нет.

Ученые в области науки о данных рассматривают аналитиков как художников в области данных. Это именно те люди, которым предстоит разобраться в корпоративных данных. Они смогут создать из них нечто

такое, что поможет решать проблемы элегантно и эффективным способом. Подобно тому как художник превращает краски в картину, которую можно повесить в доме, художники в области данных превращают данные в решение бизнес-задачи.

Выдающиеся профессионалы в области аналитики — столь же художники, сколь ученые. Именно это сочетание делает их выдающимися. Если вы скептически относитесь к этому утверждению, спросите своих знакомых аналитиков, что еще они умеют, чем увлекаются. Вы удивитесь, узнав, сколь многие из них также обладают талантом в области музыки, искусства или других творческих дисциплинах.

Нужна ли аналитику сертификация?

В последнее время часто обсуждается вопрос о разработке программ сертификации для профессионалов в области аналитики. Подобные программы используются для сертификации бухгалтеров или специалистов по финансовому планированию. Есть ли необходимость в программе сертификации для аналитиков, которая позволяет работодателям определить, кто из соискателей удовлетворяет минимальному набору критериев?

Я участвовал в дискуссиях об этих программах. Трудность состоит в том, чтобы точно определить то, что необходимо протестировать. Техническая подготовка, которую легко проверить, — всего лишь начальная ставка, когда дело касается поиска выдающегося профессионала в области аналитики. Несложно проверить, умеет ли человек кодировать или выдвигать предположения на основе линейной регрессии. А как оценить творчество? Как протестировать интуицию или деловую смекалку? Как проверить навыки презентации? И как оценить все эти характеристики в контексте аналитики? Это очень трудно.

Это хорошо, если специалист демонстрирует наличие навыков и мотивации, сдавая соответствующий экзамен. Проблема заключается в том, что любая сертификация, разработка которой целесообразна с точки зрения стоимости и обоснованности, в значительной мере будет сосредоточена на технических навыках. Такой экзамен может стать одним из первоначальных критериев и по крайней мере докажет, что человек имеет достаточную техническую подготовку и мотивацию, чтобы получить сертификат. После этого вы сможете выяснить, обладает ли он другими качествами, например творчеством, которые действительно нужны как дополнение к техническим навыкам. При таком

использовании программа сертификации может быть очень полезной. Однако как самостоятельный эталон или стандарт программа сертификации, скорее всего, не работает.

Будет ли программа сертификации принята сообществом аналитиков? Конечно, если различные организации пытаются разработать такие программы, то рано или поздно обозначатся самые эффективные из них. Тем не менее, какими бы хорошими ни были экзамены, не стоит нанимать людей, основываясь исключительно на уровне их технической подготовки. Организации могут даже не требовать наличия сертификата. Факт сдачи такого экзамена будет иметь значение только в соответствующем контексте.

В конце концов, хорошие специалисты просто понимают суть. Они понимают данные. Они знают, как их использовать, организовать, и они могут увидеть в них закономерности. Выдающиеся профессионалы в области аналитики умеют вникнуть в бизнес-задачу. Они понимают, почему то, о чем их попросили, важно и почему поставленная задача должна быть решена. Они понимают, какие существуют ограничения и как предоставить результаты, отвечающие потребностям тех, для кого производится анализ. Профессионалы в области аналитики умеют правильно формулировать проблему. Что более важно — выручка или прибыль? В чем действительно заключается суть проблемы и почему? Как следует разработать аналитический процесс? И, наконец, хорошие профессионалы в области аналитики считают себя не только учеными, но и художниками!

Обзор главы

Самые важные уроки этой главы.

- ▶ Принимая на работу аналитиков, техническую подготовку и образование рассматривайте в качестве отправной, а не конечной точки.
- ▶ Рассмотрите вопрос о найме специалистов с опытом работы в других отраслях, поскольку в них есть многое из того, чему стоит поучиться.
- ▶ Ответственность, творчество, деловая смекалка, навыки презентации и интуиция — крайне важные и часто недооцененные характеристики профессионального аналитика.

- Далеко не все, кто имеет необходимые технические навыки, обладают перечисленными в предыдущем пункте качествами.
- Выдающиеся специалисты в области аналитики сосредоточиваются на улучшении бизнеса, а не на достижении совершенства. Очень важно знать, когда результаты позволяют принять решение и можно переходить к следующей проблеме.
- Хорошие профессионалы в области аналитики связывают уровень точности данных с необходимым уровнем детализации решения. Несовершенные данные могут быть достаточными, чтобы эффективно ответить на множество вопросов.
- Сегодня услуги офшорных аналитиков сосредоточены практически исключительно на технических навыках. Пользуйтесь ими только при наличии хороших штатных специалистов.
- Хотя иметь точные результаты очень важно, не менее 50% успеха проекта зависит от того, как аналитик умеет представлять и позиционировать свои результаты спонсорам проекта, далеким от технических вопросов.
- Многие организации занимаются разработкой программ сертификации аналитиков. Время покажет, получат ли эти экзамены широкое применение. В любом случае они будут только отправной точкой в процессе оценки кандидата.
- Многих людей удивляет тот факт, что самые лучшие специалисты — не только ученые, но и художники в области данных. Не стоит недооценивать важность художественного таланта в профессиональном аналитике.

Что такое хорошая аналитическая команда?

Многие организации испытывают трудности со структурированием аналитических команд. В отличие от кадрового и финансового отделов, для аналитиков нет определенного места или даже четкой сферы деятельности. Напоминаю, что, когда мы говорим об аналитике в этой книге, мы имеем в виду такие вещи, как прогностическое моделирование, интеллектуальный анализ данных и другие методы углубленной аналитики, а не отчетность и создание электронных таблиц. Когда мы говорим о профессионалах в области аналитики, мы подразумеваем людей, которые выполняют такую работу. Многие компании имеют в своем штате специалистов, работающих в различных подразделениях, где решаемые проблемы, используемые методы и даже необходимая для работы подготовка сильно варьируются.

В случае с другими отделами дело не всегда обстоит так же сложно. Кадровый отдел, например, представляет собой централизованную организацию. Даже если в нем работают люди, которые занимаются подбором персонала для различных бизнес-единиц, описание должности специалиста по кадрам, его обязанностей и необходимых навыков довольно четкое. С аналитиками все иначе. Подумайте о разнице между аналитическими задачами, которые решают оперативный отдел и отдел мерчандайзинга. Или о том, на чем сосредоточивают свое внимание команда по управлению рисками и маркетинговая команда.

В связи с этим возникает несколько вопросов. Какую структуру следует выбрать организации для аналитической команды? Как эти команды вписываются в общую структуру организации? Какой вариант

позволит организации добиться наибольшего успеха? Какие вопросы должны быть решены перед началом работы? В этой главе мы коснемся некоторых общих проблем, возникающих перед аналитическими командами, независимо от их места в организационной структуре. Эти проблемы необходимо решить, если вы хотите создать отличную команду аналитиков, которая сможет укротить большие данные. Давайте разбираться!

Все отрасли разные

В ряде отраслей — это банковское дело, финансы и логистика — аналитика встроена в процессы принятия решений. Компании имеют в своем штате множество аналитиков. Управление рисками, например, основано на аналитике. Все предложения кредитных карт, которые вы получаете по почте, генерируются на основе аналитики. Прежде чем предложение оформить карту попадет в ваш почтовый ящик, ваши данные будут исчерпывающе проанализированы, чтобы определить уровень риска, связанный с выдачей вам кредита. Очень сложно найти банк, в котором не было бы хороших аналитиков, за исключением разве что небольшого регионального банка. В сфере предложения кредитов аналитика настолько сильно укоренилась, что никто не осмелится действовать без анализа. Целые компании добились огромных успехов благодаря тому, что они лучше всех остальных умеют выбирать надежных клиентов, которые не откажутся от выплаты долга.

В других отраслях дела обстоят не столь однозначно. Существуют компании, широко применяющие аналитику, и такие, кто практически этого не делает. Примерами могут служить розничная торговля и производство. Есть изошренные в использовании аналитики производители. Но есть и такие, кто ограничивается простым анализом электронных таблиц. Насколько более успешными могли бы быть эти производители, если бы уделяли анализу больше внимания?

Существуют известные розничные сети, которые посылают одежду одних и тех же размеров во все магазины. Вы замечали, что во многих магазинах никогда не бывает вашего размера? Многие розничные торговцы по-прежнему даже не пытаются согласовать ассортимент магазинов с местным спросом. Даже если общий объем их продаж соответствует нужным пропорциям размеров, существует огромный разброс в зависимости от местоположения магазина, а также от того, кто живет

рядом и совершает в данном магазине покупки. Профилирование размеров — это процесс определения правильного сочетания размеров для каждого магазина. Как правило, исходя из этого наполняется ассортимент. Одни магазины требуют наличия большего количества больших размеров, а другие — маленьких. Многие розничные торговцы уже используют этот подход, но далеко не все.

Знайте, к какой категории относится ваша отрасль

Есть хорошие и плохие новости. Если вы работаете в отрасли, в которой широко применяется аналитика, то сможете выбрать из множества специалистов, имеющих подходящий опыт. Тем не менее вашей организации придется напрягать силы, чтобы не отставать от лидеров. Если ваша организация относится к отрасли, в которой аналитика еще недостаточно укоренилась, то у вас будет возможность опередить конкурентов. Однако в этом случае у вас не будет столько же проверенных формул, которые можно было бы использовать.

Даже в рамках одной компании аналитика не всегда используется последовательно. В одних областях бизнеса компания может быть весьма изощренной, а в других — нет. Например, телекоммуникационная компания может очень эффективно использовать аналитику в маркетинговых целях для предупреждения оттока клиентов к конкурентам, но при этом быть совершенно неискушенной в создании моделей для прогнозирования эффектов ценообразования.

Организации предстоит решить множество проблем, прежде чем ее аналитическая команда начнет работать. Если эта команда уже существует, проблемы будут возникать по мере ее роста. Это касается всех отраслей. Если вы читаете эту книгу, то я предполагаю, что вы хотите помочь своей организации более эффективно использовать аналитику в целом, а также укротить с ее помощью большие данные. Если я прав, то вам необходима отличная аналитическая команда!

Просто начните действовать!

Важно не останавливаться в состоянии нерешительности только потому, что вы не знаете, с чего начать. Это самое худшее из того, что может произойти, поскольку приводит к пустой трате времени и мешает развитию; так вы упустите преимущества. Если вы наняли нужных людей, о чем мы говорили в главе 8, то они помогут найти лучший способ организовать команду для обеспечения максимальной эффективности.

Поручите правильным людям решение правильных задач. Если вы сможете это сделать, то внести изменения в организационную структуру не составит труда.

Несколько лет назад компания, с которой я работал, решила начать применять аналитику. Мы создали первые маркетинговые модели, а позднее помогли этой компании обновить методы измерения, используемые для рекламных акций. Это было очень успешное сотрудничество. На протяжении нескольких лет эта компания продолжала обращаться к нам.

Совместно проведенная нами аналитическая работа помогла компании осуществить ряд важных изменений. Сотрудники компании определили некорректно оцененные сегменты клиентов, к которым они обращались с неподходящими маркетинговыми предложениями. Они это исправили и разработали единую точку зрения на деятельность, связанную с продвижением, вместо нескольких, используемых в различных частях организации. Единая концепция позволила легко договориться о том, во что следует инвестировать и что наиболее эффективно. Отделу маркетинга больше не нужно было спорить с финансовым отделом. Они также смогли сделать свой бизнес более клиентоориентированным, добавив метрики, касающиеся клиентов, в широкий спектр отчетов и аналитических процессов.

Спустя несколько лет эта компания начала создавать собственную команду штатных аналитиков и взяла на себя обязательство инвестировать в аналитику в долгосрочной перспективе. Мы помогли им получить первые результаты, однако они могли бы реализовать больше преимуществ более быстрыми темпами, если бы начали действовать раньше. Они так беспокоились о том, с чего начать, что затянули процесс на долгое время и не воспользовались всеми возможными преимуществами.

Дефицит талантов

По мере становления аналитической команды в вашей организации нужных людей будет не хватать. Потребуется по крайней мере несколько отличных специалистов, а таких людей очень немного. Специалистов вообще-то мало во многих отраслях, но в аналитике — особенно. Отчасти это объясняется двумя причинами. Во-первых, спрос на профессионалов в области аналитики стремительно растет. О быстром увеличении спроса на талантливых аналитиков свидетельствуют

многочисленные блоги, статьи и книги, включая ту, которую вы читаете. Во-вторых, талантливых аналитиков система образования всегда выпускала немного. Адаптация академических программ для повышения качества подготовки специалистов требует определенного времени.

Компании вынуждены конкурировать между собой за талантливых сотрудников. Недостаточно просто предложить большую зарплату или дополнительные выгоды. Необходимо убедиться, что аналитикам предлагают сложные задачи и их деятельность получает поддержку. Если штатные аналитики, независимо от того, сколько им платят, понимают, что их работодатель несерьезно относится к аналитике и что они не оказывают влияния, то они уйдут. Профессионалам в области аналитики нужны деньги, как и всем остальным, но, как любому другому человеку, не менее важно признание, возможность оказывать влияние и улучшать свои навыки.

Будьте скромны

Во время экономического спада нетрудно предположить, что все отчаянно нуждаются в работе, однако талантливые аналитики пользуются спросом. Если вы полагаете, что таких людей легко заменить, вы их потеряете. Вам может быть трудно добиться повышения зарплаты для нового сотрудника в сложных экономических условиях. Однако если это все, что нужно, чтобы заполучить подходящего человека, то это небольшая цена и хорошая инвестиция.

Даже в тяжелых экономических условиях 2009–2011 годов профессионалы в области аналитики могли найти хорошие рабочие места. Специалисты по подбору персонала часто объявляют о вакансиях, а участники профессионального сообщества обсуждают условия работы, о которых они слышали. Профессионалы в области аналитики пользуются таким спросом, что в одной из худших экономических ситуаций столетия все еще слышат о вакансиях и отвечают на запросы специалистов по подбору персонала. Если в тяжелых экономических условиях ваша компания готова пойти на снижение уровня заработной платы вместо того, чтобы вести переговоры, то вы должны убедить свой кадровый отдел немного ослабить эти требования, чтобы найти талантливых специалистов в области аналитики. Это продлится ближайшие несколько лет: пока еще мало аналитиков, имеющих опыт работы с большими данными. Учитывая многообразие новых источников больших данных и новых инструментов, используемых для их укрощения, специалистов, обладающих нужным опытом, будет трудно отыскать.

Структуры команд

Как следует организовать аналитические ресурсы, чтобы все бизнес-единицы, нуждающиеся в аналитической поддержке, могли получить необходимое и в то же время сохранить согласованность структуры предприятия? Ответ на этот вопрос пытались найти многие авторы, в том числе Том Дэвенпорт и Джоан Харрис¹⁷. Здесь мы обобщим основные варианты структур, случаи, когда их использование уместно, и принцип их работы. Обратите внимание: аналогичные структуры применимы и к другим группам в рамках компании, однако мы сосредоточимся на их применении в области аналитики. Основные структуры разделены на три категории: децентрализованные/функциональные, централизованные и гибридные. Выбор наилучшего варианта для конкретной организации может быть непростым.

Недавно одна компания, работающая в индустрии развлечений, решила значительно повысить эффективность использования аналитики. В компании существовали независимые подразделения. Поскольку многие из них были приобретены, их независимость сохранялась намеренно, так как каждое выполняло свою работу, имело свой стиль и культуру.

Одно из подразделений решило заняться прогнозной аналитикой. Головная организация положительно отнеслась к этой инициативе, однако другие подразделения не проявили к этому никакого интереса. Возник вопрос: следует ли головной организации позволить подразделению работать самостоятельно и надеяться, что в будущем остальные бизнес-единицы начнут применять такой же подход? Подойдет ли метод, выбранный данным подразделением, для других подразделений? Следует ли корпоративной команде разработать план и позволить данному подразделению стать первым, кто будет его применять? Со временем остальные подразделения могли бы использовать те же «официальные» процессы.

На эти вопросы нельзя ответить однозначно. Можно позволить подразделению поработать и добиться некоторых успехов. Затем компания при необходимости может подстроить разработанные процессы под другие подразделения. Можно посоветовать с самого начала разработать нечто более универсальное с корпоративной точки зрения. Правильный ответ будет зависеть от культуры организации и от того, что ей кажется уместным. Компания в данном случае решила выбрать

нечто среднее: она позволила подразделению взять на себя инициативу, а головной организации — принять участие в процессе.

Децентрализованные/функциональные структуры

В децентрализованной организации аналитические команды подотчетны специальному функциональному подразделению. Команда аналитиков отчитывается перед группой, которую поддерживает. При использовании этой модели специалисты, занимающиеся операционной аналитикой, подотчетны операционной команде и главному операционному директору (chief operations officer — COO). Маркетинговые аналитики подотчетны маркетинговой команде и главному директору по маркетингу (chief marketing officer — CMO). Риск-аналитики подотчетны команде, занимающейся управлением рисками, и т. д.

Преимущество такой структуры состоит в том, что специалисты по анализу работают именно там, где они нужны. Каждый день они видят людей, которых поддерживают, и проблемы, которые им необходимо решать. Как правило, организации начинают именно с этой модели потому, что какая-либо часть организации начинает применять аналитику раньше остальных. В результате этой бизнес-единице первой требуется нанимать профессиональных аналитиков. Вот почему организации чаще всего начинают с использования децентрализованной, функциональной модели. В самом начале децентрализованная модель подразумевает наличие одной небольшой команды аналитиков, подчиняющейся одной бизнес-единице.

Недостаток децентрализованной модели состоит в том, что ресурсы в конечном счете оказываются разбросанными по организации. Многие сотрудники с одинаковой подготовкой и навыками не входят в одну и ту же структуру. У них нет никакой формальной, а иногда и неформальной связи. Каждая аналитическая команда работает только с функциональным подразделением, в составе которого находится. Эта ситуация неидеальна с точки зрения долгосрочной перспективы. В данном случае одна команда может не иметь возможности в случае необходимости «одолжить» людей у другой, даже если другая команда располагает избыточными ресурсами.

Потенциальной проблемой, связанной с децентрализацией, является также отсутствие четкого карьерного пути. Рассмотрим организацию, состоящую из пяти бизнес-единиц, в каждой из которых есть три

или четыре штатных аналитика. В рамках данных бизнес-единиц у них не так уж много возможностей для продвижения. В лучшем случае они смогут возглавить команду, состоящую из трех-четырех человек. И то только после ухода существующего начальника. Это не очень привлекательная карьера.

В этом примере в рамках всей организации числится около двадцати аналитиков. Ни у кого из них нет достаточно возможностей для продвижения по службе, и большинство не поддерживают никаких контактов со специалистами за пределами своих подразделений. Таким образом, децентрализация в чистом виде может применяться в качестве краткосрочной или среднесрочной меры на начальных этапах внедрения аналитики. В долгосрочной перспективе следует подумать либо о централизованной, либо о гибридной модели.

Это не означает, что децентрализованные команды долго нигде не могут использоваться. В индустрии авиаперевозок, например, одна команда аналитиков может заниматься управлением выручкой, а другая — маркетингом. Типы проводимого анализа, необходимые инструменты и навыки могут быть настолько разными, что это затруднит эффективное объединение команд. Это нормально. Просто периодически анализируйте и другие варианты, чтобы убедиться, что в изменении структуры нет необходимости.

Централизованные структуры

Централизованная структура подразумевает наличие одной главной аналитической команды, занимающей определенное место в организационной структуре. Эта команда поддерживает все бизнес-единицы и обеспечивает их аналитические потребности. Одной из проблем в данном случае будет определение места централизованной команды. Централизованная аналитическая команда может быть подотчетна главному финансовому директору (chief financial officer — CFO), главному операционному директору (COO), главному директору по информационным технологиям (chief information officer — CIO). Для централизованной команды аналитиков не существует установленного места в структуре организации, поэтому каждая организация поступает по-своему.

Преимущество централизованной команды — в возможности перераспределения ресурсов по мере необходимости. Рассмотрим

организацию, в операционную и маркетинговую команды которой входят по три аналитика. Операционная команда переживает спад рабочей нагрузки и не нуждается в частом проведении анализа или в освоении характерного для данной команды бюджета. В то же время маркетинговая команда перегружена работой в связи с внедрением некоторых новых крупных инициатив. Децентрализованная модель не позволяет аналитикам из операционной команды помочь отделу маркетинга. А при использовании централизованной структуры человек, управляющий всеми аналитиками, может легко перемещать их в рамках организации. Централизованная структура помогает снизить риски, связанные с изменением спроса во времени.

Другое большое преимущество централизованной структуры состоит в том, что у аналитиков появляется возможность получить опыт работы в различных подразделениях компании и освоить различные типы анализа. Хорошим специалистам будет скучно делать одно и то же в течение десяти лет. Однако если за эти же десять лет человек поработает в различных отделах, освоит целый ряд новых методов и познакомится с множеством новых людей, то это будет замечательно. Такие условия способствуют повышению профессионализма аналитической команды. При этом выигрывают и специалисты, и организация в целом.

Централизованная, но преданная делу

Даже если ваша организация выберет полностью централизованную структуру, рассмотрите возможность сделать так, чтобы ресурсы были ориентированы прежде всего на конкретные бизнес-единицы. Работа с одними и теми же профессионалами позволит бизнес-командам быстрее освоиться в использовании аналитики. Не стоит недооценивать отношения, налаженные со специалистами, с которыми они работают на постоянной основе.

К недостаткам централизованных команд можно отнести вероятность того, что в организации в конечном счете окажется множество специалистов широкого профиля и ни одного глубоко разбирающегося в конкретной области. На работе бизнес-единицы негативно сказывается частая смена работающих с ней аналитиков. По этой причине даже полностью централизованная команда может поручить конкретным людям помочь конкретным бизнес-единицам. Официально аналитики подотчетны централизованной группе, но на практике внедряются в бизнес-единицу, на работу с которой они назначены.

С точки зрения повседневной работы аналитик считается частью данной команды.

Централизованная аналитическая команда часто выставляет бизнес-единицам счет за свои услуги. Однако иногда оплата работы этой команды считается корпоративными накладными расходами. Если команда выставляет счет подразделению, это позволяет удостовериться в том, что люди занимаются действительно важной работой, и тогда бизнес-единицы вынуждены расставлять приоритеты. В то же время, когда за услуги специалистов должно платить определенное подразделение, труднее использовать инновационные аналитические методы. В идеале организации выделяют определенный корпоративный бюджет, чтобы аналитическая команда могла работать над инновационными проектами сверх того, что будут оплачивать отдельные бизнес-единицы. Укрощение больших данных — отличная цель для такого бюджета. Спонсируйте ранние стадии работы на корпоративном уровне, а затем позвольте подразделениям оплачивать дальнейшую работу, когда они осознают значение больших данных.

Не стоит недооценивать важность такого структурирования организации, чтобы обеспечивалось использование инновационной аналитики. Не следует ожидать, что конкретные бизнес-единицы с ограниченными бюджетами смогут в полном объеме финансировать инновационные проекты. Инновационная аналитика должна спонсироваться и поддерживаться на высшем уровне и рассматриваться в качестве стратегической корпоративной инвестиции.

Гибридные структуры

Гибридные структуры представляют собой именно то, что следует из их названия: наличие централизованной команды, а также специализированных групп в конкретных подразделениях. Такой тип структуры может возникнуть по разным причинам. Часто это происходит, когда какая-то одна бизнес-единица возглавляет процесс аналитики. Например, это подразделение создало сплоченную команду аналитиков и не готово отказаться от контроля над ней. В то же время другие подразделения начали более интенсивно применять аналитику. Для поддержки других подразделений формируется централизованная команда. Однако первоначальная команда остается в подразделении, где она была создана.

Другая распространенная гибридная модель подразумевает существование главной централизованной команды, которую часто называют центром передового опыта (center of excellence — COE), или экспертным центром (center of expertise — COE). Задача специалистов данной команды состоит в поддержке предприятия в целом. Хотя большинство аналитиков работают в конкретных подразделениях, существуют свободно перемещающиеся между бизнес-единицами сотрудники, призванные поддерживать согласованность используемых подходов и инструментов. Команда COE также аккумулирует новые знания, получаемые аналитиками в процессе своей работы в различных подразделениях. Команды аналитиков, работающие в бизнес-единицах, могут быть либо формально, либо неформально подотчетны главной группе аналитиков.

Не беспокойтесь о структуре — беспокойтесь о людях

Следует еще раз отметить, что самое главное заключается не в том, как вы структурируете свои аналитические команды. Самое главное — это то, что у вас есть правильные люди, производящие правильный анализ исходя из правильных причин. Кроме того, важно сосредоточиться на создании среды и культуры, позволяющих вашей организации нанимать, развивать и удерживать талантливых аналитиков.

Поддержание высокого уровня компетентности команды

В аналитической команде работают люди с различным уровнем профессионализма, как и в любой другой команде: один-два опытных профессионала наряду с новыми сотрудниками, недавно окончившими обучение. На начальном этапе необходимо сосредоточить внимание на конкретных навыках, которые крайне необходимы в данный момент для решения существующих проблем.

По мере того как команда будет расти, важно вводить туда людей, обладающих различными аналитическими навыками. Если первые нанятые вами специалисты имеют сильную подготовку в области интеллектуального анализа данных, далее имеет смысл искать людей с опытом работы в сфере оптимизации или прогнозирования. Участие специалистов с опытом работы в различных областях аналитики

предоставляет больше возможностей для выявления новых способов повышения ценности бизнеса. Кроме того, по мере роста команды можно сосредоточиться на создании карьерных возможностей. Изначально нанятые люди, как правило, достаточно опытные, поскольку они самостоятельно и успешно работают, практически без помощи других экспертов. По мере роста команды можно начинать нанимать менее опытных аналитиков и развивать их.

Матричный подход

Матричный подход — неиерархический подход к управлению деятельностью команды — помогает сохранить высокий уровень навыков аналитиков. Он предполагает назначение руководителя конкретного проекта, которому поручается выполнение нескольких ключевых функций. Во-первых, он отвечает за управление проектом. К счастью, как правило, в аналитических проектах объем работы, связанной с управлением им, невелик. Во-вторых, руководитель проекта отвечает за определение направления проекта, разработку плана анализа и соблюдение сроков. Третья и самая важная функция предусматривает обобщение результатов, их интерпретацию и разработку рекомендаций, а также подготовку представления результатов работ по проекту. Под началом руководителя аналитического проекта находится один или несколько специалистов, отвечающих за выполнение работ по проекту.

В команде не обязательно должен присутствовать человек, которого называют руководителем аналитического проекта, поскольку дело не в названии, а в выполнении необходимых для конкретного проекта функций. Например, в команде есть два специалиста — Боб и Сью. При работе над проектом А Сью может быть руководителем, а Боб — ее подчиненным. При работе над проектом Б руководителем может быть Боб, а Сью — работать под его началом. Руководителем выбирают того, кто лучше всего подходит для выполнения соответствующих функций. Например, если проект в значительной степени ориентирован на прогнозирование, то руководителем будет выбран человек, который хорошо знаком с данной областью.

Разумеется, самых сильных и опытных аналитиков чаще выбирают руководителями проектов, а новым и не слишком опытным специалистам чаще отводится роль подчиненных. Однако назначение функций исключительно на основе старшинства или стажа может привести

к путанице. Вот почему использование матричного подхода для управления командой аналитиков — хорошая идея. Когда члены команды обмениваются должностями, это развивает сплоченность. Каждый человек старается держать свое эго под контролем, поскольку знает, что время от времени ему придется работать под началом одного из коллег. Члены команды могут по-настоящему узнать друг друга и проявить свои самые сильные стороны, а также многому друг у друга научиться, что подводит нас к теме взаимного обучения.

Войдите в матрицу!

Матричный подход к управлению командой аналитиков помогает увеличить производительность труда, сплотить команду и обеспечить рост компетентности специалистов. Сосредоточившись на способностях людей, а не на их званиях или должностях, вы создадите культуру, в которой люди имеют больше шансов добиться успеха и где все сосредоточено на получении правильных ответов, а не на том, кто именно их находит.

Взаимное обучение

Один из самых важных аспектов работы аналитической команды независимо от ее структуры — обеспечение возможности взаимного обучения специалистов. Если один из них выдающийся программист, убедитесь, что он делится с коллегами своим опытом или создает письменное руководство, где изложены советы и тонкости работы, а также наставляет других или даже проводит небольшие уроки. Члены команды захотят совершенствоваться. Одним из лучших способов является освоение новых для них областей аналитики, а также совместная работа над проектами. Такое обучение будет полезно как обучающему, так и обучаемому.

Менеджерам нельзя терять хватку

Менеджеры и руководители должны участвовать в рабочем процессе и не терять своих навыков. Люди ненавидят «пустые костюмы», которые занимают высокое положение в организационной структуре, но не знают, что и как следует делать. Такой «пустой костюм», возможно, когда-то много знал, но со временем потерял свои навыки. В любом случае считается, что такой человек мало что умеет. Он, может быть, хорошо говорит, но не делает то, о чем говорит.

Профессионалы в области аналитики часто даже более чувствительны к таким вещам, чем представители других профессий. То же самое относится ко многим техническим областям. Технически подкованные люди, как правило, не уважают руководителей, которые указывают им, что делать, стоят над душой и критикуют их работу, если эти руководители не являются специалистами в том, о чем говорят. Если аналитик видит, что руководство не разбирается в том, о чем говорит, то сохранить его уважение будет практически невозможно. Это не означает, что человек, не обладающий глубокими познаниями во всех областях аналитики, не может управлять аналитиками. Менеджеру важно признать границы своих знаний и довериться команде в деталях, выходящих за рамки его опыта.

Навыки менеджера должны оставаться на высоком уровне... как у рыцаря-джедая!

В сериале «Звездные войны» Йода часто руководил действием, находясь за кадром. Однако при необходимости он вступал в бой и сражался наравне с лучшими воинами. Точно так же хороший менеджер может не каждый день находиться «в поле», но, когда потребуются, он должен быть в состоянии производить анализ наравне с лучшими специалистами. Команда будет всегда оставаться в тонусе, если знает, что при необходимости лидер сможет лично позаботиться о деле.

Для того чтобы менеджеры не утратили свои навыки, подумайте о том, чтобы их обязанности менялись по крайней мере один раз в год. Пусть они поработают «в поле» и позанимаются анализом. Это отличный способ сохранить навыки актуальными, хотя реализация этого плана и представляет определенные трудности. Существуют компании, которые используют такой подход. Я знаю сеть ресторанов, в которой каждого сотрудника обязывают работать в ресторане каждый год по нескольку дней, чтобы все понимали, что там происходит. Это позволяет всем сотрудникам сохранять контакт с реальным миром, и мои знакомые, работающие там, находят такое ежегодное упражнение очень ценным.

Кто должен заниматься углубленной аналитикой?

На форумах сообщества аналитиков часто затрагивается одна тема. Я даже обсуждал ее в моем блоге¹⁸. Вопрос заключается в следующем. Учитывая существующее в настоящее время программное обеспечение

и инструменты с дружелюбным к пользователю интерфейсом, которые дают возможность заниматься углубленной аналитикой, стоит ли позволять людям, не имеющим соответствующей подготовки, самостоятельно производить анализ, используя эти инструменты?

Дело в том, что под «простым интерфейсом» инструмента нередко подразумевается возможность его правильного использования. Однако (см. главу 6) это не так. То, что инструмент легкий в использовании, совсем не означает, что его легко использовать правильно. Именно простота использования позволяет быстро и неосознанно делать именно неправильные вещи. Например, инструменты, генерирующие код SQL с помощью интерфейса point-and-click, дают возможность объединять данные так, как пользователям заблагорассудится. Эти средства используют необходимый синтаксис, но не гарантируют, что этот синтаксис будет иметь хоть какое-то значение.

Организация должна убедиться, что человек, который будет применять тот или иной инструмент, имеет подходящие навыки и опыт. Аналитический инструмент может снять с пользователя часть нагрузки, связанной с программированием, однако пользователю необходимо разбираться в генерируемых результатах. Если бы вы были уверены в том, что новичок может задавать правильные вопросы, что все необходимые данные подготовлены и доступны в нужном формате и что точно известно, какой алгоритм следует применять, то добиться успеха мог бы практически любой человек. В таком случае пользователь действительно мог бы просто нажимать на кнопки. В реальном мире так не бывает.

Многое зависит от создания подходящего аналитического процесса или модели, а это уже выходит за рамки простого использования интерфейса инструмента. Правильно ли было предсказано поведение? Поддерживает ли это предсказание самый лучший набор независимых переменных? Достаточно ли у аналитика опыта, чтобы понять, что возникли проблемы? Знает ли он, как их решать? Мы говорили в главе 7 о том, что для проведения углубленного анализа не существует волшебной кнопки!

Это не означает, что новички, не имеющие специальной подготовки или навыков, не могут создать дополнительную ценность для организации. Это просто вопрос гарантии того, что они не превысят своих полномочий и не будут делать того, что им не следует. Большая часть сотрудников организации должны использовать заранее определенные

шаблоны или отчеты; возможно, им будет поручен некоторый дополнительный анализ. Более сложную работу следует оставить экспертам. Формальная аналитическая команда должна состоять в основном из таких экспертов.

Примеры непоследовательности

Комментарии, приведенные в предыдущем разделе, иногда наводят на мысль о том, что профессионалы в области аналитики занимают оборонительную позицию. На самом деле такие правила приняты во многих других областях деятельности. Почему-то люди, далекие от аналитики, не распространяют на нее логику, которую применили бы в иной области. Рассмотрим, почему следует позаботиться о том, чтобы углубленной аналитикой занимались подходящие люди.

Джейн решила, что она больше не хочет быть аналитиком, а желает заниматься созданием рекламы для отдела маркетинга. Она установила на свой компьютер те же современные и мощные инструменты для создания графики и генерации контента, которые используют маркетологи. Она потратила несколько часов, чтобы научиться пользоваться этими инструментами. Джейн смогла легко создавать брошюры, графику и все, что ей нужно, поскольку программное обеспечение позволяет простым щелчком кнопки мыши организовать фотографии, изображения и текст. Она идет в отдел маркетинга и говорит: «Я установила все необходимое программное обеспечение на свой компьютер и прошла обучение. Я хотела бы присоединиться к вашей команде и заниматься прямой почтовой рекламой, создавать журнальные объявления и брошюры о продуктах. Примете меня?» Сотрудники отдела маркетинга рассмеялись бы ей в лицо: для того чтобы создать хороший контент для маркетинга, мало уметь щелкать кнопками.

Джон решил, что хочет работать в команде CFO и каждый месяц заниматься закрытием отчетов. Он узнал, какое бухгалтерское программное обеспечение использует его компания. Как и Джейн, потренировался в использовании этого программного обеспечения. Затем он идет в бухгалтерию и говорит: «Я бы хотел помогать вам с закрытием отчетов каждый месяц. У меня нет никакой подготовки в области финансов или бухгалтерского дела. Однако я прошел все курсы, доступные для программного пакета, который вы используете. Я знаю, какие пункты меню необходимо выбирать на каждом этапе закрытия

отчета. Когда я могу начать работу?» Как вы думаете, получит Джон эту работу?

И последний пример: соседу Джо необходимо спилить большое дерево. Он спрашивает у Джо, какую службу тот может порекомендовать. Джо отвечает: «Тебе не нужна никакая служба. Я только что купил первоклассную бензопилу. Я прочитал руководство от корки до корки, лезвия наточены, и пила готова к работе. Я спилю твоё дерево». Какой сосед стал бы всерьез рассматривать такое смешное предложение?

Не ищите легких путей

Большинство людей посмеются, услышав о том, что новичок хочет заняться созданием рекламы для крупного маркетингового проекта или закрытием корпоративных книг; не воспримут всерьез слова соседа с первоклассной бензопилой, который предложил спилить дерево. Почему же тогда многие готовы предположить, что человек, не обладающий соответствующим опытом, но научившийся использовать аналитический инструмент, может создавать высококачественные аналитические процессы? Не попадайте в эту ловушку!

Организация, которая хочет создать отличную команду аналитиков, должна помнить, что эффективная аналитика — это и наука, и искусство. Как великие художники, вероятно, не смогли бы нарисовать шедевр, впервые взяв в руки кисть, так и аналитическая команда не может стать выдающейся в первый же день работы. Это приходит с практикой и опытом. Как и в других дисциплинах, в аналитике существуют сложности и нюансы, которые неочевидны для тех, кто не знаком с этой областью. Вы не сможете позволить новичку без необходимых навыков, образования и опыта делать вашу собственную работу? Тогда не стоит позволять новичку без необходимых навыков, образования и опыта создавать передовые аналитические процессы. Как говорилось в главе 8, крайне важно, чтобы аналитическая команда состояла из профессионалов, а не просто из людей, которые перешли из других отделов организации и получили новую роль и должность.

Создайте такие условия, чтобы новички могли достичь успеха

Теперь посмотрим на эту тему с более позитивной точки зрения. В организации многие сотрудники хотят использовать средства углубленной аналитики. Предположим, Барб из отдела маркетинга является одной из них и она готова взять на себя ответственность за аналитику в своей

бизнес-единице. Она готова упорно работать. Это очень хорошо, и аналитическая команда, конечно, поддержит Барб и поможет ей. Однако начинать следует не с установки программного обеспечения на ее компьютер, чтобы она могла начать самостоятельно проводить анализ.

Вернемся к тому, что дружественные к пользователю аналитические инструменты способствуют повышению производительности труда. Возражение против того, чтобы инструменты были доверены неподготовленным людям, продиктовано вовсе не желанием защитить рабочие места профессиональных аналитиков и запретить использование новых технологий. Если бы дело было в этом, то дружественных к пользователю аналитических инструментов просто не существовало бы.

Человек с ограниченным видением способен утверждать: если он является частью команды из десяти специалистов, а использование инструмента позволяет сократить время работы аналитика в два раза, то это означает, что половину команды можно уволить. Любой, кто так думает, должен покинуть команду, поскольку никогда не сможет стать выдающимся аналитиком. Правильно было бы рассматривать данную ситуацию так: эти десять человек оправдывают свое существование тем, что они делают сегодня. Если вдруг возникнет возможность выполнять текущую работу в два раза быстрее, значит, половину своего времени они смогут потратить на решение новых проблем и создать еще большую дополнительную ценность. Современные удобные инструменты еще больше будут оправдывать существование этих специалистов, помогать им в развитии своего мастерства и ставить перед ними дополнительные задачи. Эта ситуация является выигрышной с любой точки зрения.

Пусть каждый занимается своим делом

Если специалисты в области бизнеса, далекие от аналитики, захотят ею воспользоваться, им не обязательно браться за сложную работу. Работа аналитической команды заключается в том, чтобы аналитика применялась по всей организации. Пусть бизнесмены тратят свое время на продвижение идеи использования аналитики для изменения бизнес-процессов, которыми они управляют. Если аналитические команды будут делать то, что они умеют делать лучше всего, а бизнес-команды — то, что *они* делают лучше всего, то выиграют все.

Надеюсь, вы убедились, что дружественные к пользователю аналитические технологии сами по себе являются хорошей вещью. Как же организации следует использовать их? Главное заключается в том, чтобы помочь Барб достичь своих целей. Это не значит, что ей следует

поручить всю сложную работу. Команда аналитиков должна работать вместе с ней, помочь ей произвести необходимый анализ и позволить делать отчеты о работе созданных ими моделей. Результаты, касающиеся клиентов, должны отображаться в программе CRM. Они должны быть доступны для любых других корпоративных приложений. Барб необходимо предоставить инструменты для использования аналитики, которая была для нее разработана. Самостоятельно выполнять сложную работу она, вероятно, хочет не больше, чем это ей необходимо.

Почему ИТ-специалисты и аналитики не ладят между собой?

Создавая аналитическую команду, организация сталкивается с серьезной проблемой. Это война между аналитиками и ИТ-специалистами. Во многих организациях между ними существует длительная вражда. Верите или нет, но для возникновения этой напряженности в прошлом действительно существовали логичные, реальные причины. Однако сегодня они неактуальны. Чтобы понять почему, рассмотрим роли, которые организации отводят ИТ-специалистам и аналитикам.

Аналитику поручено расширение границ использования корпоративных данных (табл. 9.1). Он должен придумывать инновационные методы и при этом не ограничиваться определенными рамками. В то же время ИТ-специалисты должны поддерживать работу систем и обеспечивать, чтобы каждый мог сделать то, что ему нужно. ИТ-специалистам нужно гарантировать эффективное распределение ресурсов и сохранение контроля.

Табл. 9.1

Роли аналитиков и ИТ-специалистов

Функции аналитической команды	Функции ИТ-команды
Интенсивное использование системных ресурсов	Ограничение использования ресурсов
Создание таблиц и использование большого количества пространства	Ограничение создания таблиц и использования пространства
Выполнение сложных запросов	Минимизация использования сложных запросов

Окончание табл. 9.1

Функции аналитической команды	Функции ИТ-команды
Мышление, не ограниченное определенными рамками	Ограничение пользователей определенными рамками
Экспериментирование с новыми подходами	Использование утвержденных подходов
Работа с небольшим количеством правил и ограничений	Обеспечение соблюдения правил и ограничений

ИТ-специалисты и аналитики не ладят отчасти потому, что выполняют функции, противоречащие друг другу. Одна и та же компания платит им за прямо противоположные вещи! Одной команде поручено скрывать данные, управлять ими и контролировать использование ресурсов. Другой — обрабатывать данные, используя большое количество ресурсов, и находить новые способы ведения дел. В такой ситуации возникновение конфликта практически неизбежно.

Еще более усложняет дело то, что аналитики обычно подотчетны главному директору по управлению бизнесом, а ИТ-специалисты — главному директору по информационным технологиям. Единственным человеком в организации, который имеет непосредственную власть над обеими командами, является CEO. Однако оба руководителя высшего звена вряд ли захотят привлекать CEO к разрешению разногласий между своими командами.

ИТ-специалистам и аналитикам нужно подписать мирный договор!

Очень важно, чтобы ваша организация заставила ИТ-команду и аналитическую команду прийти к соглашению о совместной работе. Доступные сегодня технологии позволяют обеим командам мирно сосуществовать и даже помогать друг другу. Самое трудное заключается в том, чтобы заставить людей преодолеть предубеждения и справиться с враждебностью. Приложите к этому усилия, иначе вам будет очень трудно создать отличную аналитическую команду.

В результате ИТ-специалисты нередко относятся к аналитикам как к ковбоям, создающим «теневую» ИТ-среду без правил и политики. А аналитики считают ИТ-специалистов помешанными на контроле людьми, которые поставили целью сдерживание прогресса и создание препятствий.

Хорошая новость: многие причины этой вражды сегодня неактуальны. В главах 4 и 5 мы рассказали о появлении песочниц, аналитики, встроенной в базы данных, и о слиянии среды данных с аналитической

средой. Эти технологии позволяют преодолеть разрыв между ИТ-специалистами и аналитиками. Если ваша организация намерена создать выдающуюся команду аналитиков, очень важно устранить этот разрыв.

Аналитиков сложно заставить признаться, что они хотели бы достичь соглашения с ИТ-специалистами. Они не стали бы управлять отдельной системой, если бы не считали это неизбежным злом, необходимым для выполнения их работы. Почему? Потому что это часто отвлекает аналитиков от работы, которую они по-настоящему любят! Давайте разберемся.

Если команда аналитиков имеет собственную аналитическую среду, которую им приходится поддерживать, то это фактически превращает их в системных администраторов, планировщиков технического обслуживания и т. д. Аналитик создает новый аналитический процесс, который должен выполняться еженедельно. Угадайте, что происходит дальше? Ему же приходится каждую неделю следить за работой этого процесса и контролировать ее! Он же должен решать проблемы, связанные с изменением в потоках данных или другими изменениями в системе, оказывающими на этот процесс влияние.

Аналитики на самом деле не хотят этим заниматься. Более того: если аналитик создаст четыре, пять или шесть новых аналитических процессов, то очень скоро практически все его рабочее время будет поглощено наблюдением за работой этих созданных процессов. У него не останется времени на проведение нового анализа! Это довольно плачевный итог. Аналитическая команда с удовольствием передала бы функции системного администрирования, планирования, резервного копирования и т. д. ИТ-отделу, сотрудники которого получают зарплату за эту работу, любят ее и являются в ней специалистами. Так гораздо эффективнее, и все будут счастливы. Это позволит аналитикам высвободить время на совершенствование своего мастерства вместо того, чтобы тратить его на наблюдение за работой процессов.

Обзор главы

Самые важные уроки этой главы.

- ▶ Вместо того чтобы без конца размышлять над структурой команды, наймите нескольких хороших аналитиков и поручите им решение правильных проблем.

- Будьте разборчивы в том, кого вы нанимаете. Успех в большей мере зависит от людей, составляющих аналитическую команду, чем от места этой команды в организационной структуре.
- На рынке труда наблюдается дефицит талантливых аналитиков. Придется потрудиться, чтобы найти подходящих людей.
- Большинство организаций начинают с децентрализованной, функциональной структуры аналитической команды. Со временем компании часто переходят к централизованной или гибридной структуре.
- Организации должны поощрять взаимное обучение и обмен опытом между членами аналитической команды.
- Попробуйте применять матричный подход к реализации аналитических проектов. Позвольте руководить более подготовленному специалисту при работе над каждым проектом.
- Менеджеры должны поддерживать уровень своих знаний и быть способными при необходимости работать наряду с остальными специалистами.
- Простые в эксплуатации инструменты — не волшебная кнопка, позволяющая неподготовленным людям производить качественный анализ. С помощью таких инструментов неподготовленные люди легко могут делать неправильные вещи.
- Аналитические команды должны позволить своим деловым партнерам успешно применять аналитику. Пусть профессионалы в области аналитики выполняют сложную работу, а затем делают результаты доступными для специалистов в области бизнеса.
- Помогите заключению мирного соглашения между командами аналитиков и ИТ-специалистов. Никто, особенно деловые партнеры, зависящие от поддержки обеих команд, не выигрывает от их взаимной неприязни.

ЧАСТЬ IV

Объединение пройденного: аналитическая культура

Создание условий для внедрения инноваций в сфере аналитики

Что такое инновация? Согласно словарю Merriam—Webster, инновация — это:

- 1) «нововведение»;
- 2) «новая идея, метод или устройство»¹⁹.

Такое определение слова «инновация» разочаровало меня, поскольку оно не отражает той мощи, которую инновации могут обеспечить бизнесу и миру в целом. Не будь инноваций, мы до сих пор жили бы в каменном веке, поскольку именно инновации являются двигателем перемен и прогресса. К сожалению, в сегодняшней бизнес-среде инновациям не всегда уделяется должное внимание.

Очень немногие компании добиваются успеха, просто следуя за конкурентами.

Инновация — один из ключевых факторов успеха. Если организация просто наблюдает за тем, как ее конкуренты начинают работать с большими данными, то проблемы обеспечены. Особенно если эта компания злорадствует по поводу того, как ее конкурент тратит деньги на эту инициативу, в то время как сама этого не делает. Возможно, услышав о преимуществах, которые обеспечила конкуренту работа с большими данными, компания захочет делать то же самое. Это просто игра в догонялки. Усилия первопроходцев затрачиваются на получение конкурентного преимущества, в то время как усилия тех, кто следует за ними, — на то, чтобы догнать лидеров. Это не формула победы.

В этой главе мы рассмотрим некоторые основные принципы, лежащие в основе инновации. Затем применим их к миру больших данных и аналитики с помощью концепции центра аналитических инноваций. Наша цель — предоставить читателям идеи для внедрения аналитических инноваций и укрощения больших данных в своих организациях.

Компаниям необходимо больше инноваций

С 2008 по 2011 год экономическая ситуация была настолько тяжелой, что многие компании боролись за выживание. Все, что не было напрямую связано с выживанием в этом месяце, в этом квартале или в этом году, отодвигалось на задний план. Однако даже в лучшие времена очень немногие компании уделяют инновациям достаточно внимания. По иронии судьбы, одним из лучших условий для опережения конкурентов являются трудные времена. Пока конкуренты притаились, компания может инвестировать в инновации и превзойти их. В плохие времена статус-кво уже не действует, и иногда это заставляет организации наконец использовать новые идеи. Невозможно просто сидеть и смотреть, как одни и те же старые стратегии продолжают приводить к успеху. Необходимо сделать что-то новое, так почему бы не подойти к делу творчески?

Без постоянного потока новых идей, продуктов и услуг компания не сможет конкурировать в долгосрочной перспективе. Так было всегда, однако сегодня в глобальном, взаимосвязанном, быстро меняющемся мире это еще более верно, чем когда бы то ни было. Инновации являются ключом к выживанию, а инновации в области аналитики и в способах использования данных заслуживают особого внимания. Как организация будет применять аналитику для выявления закономерностей, которые можно использовать, и к каким действиям это приведет? Аналитика — наиболее важный инструмент, который позволяет добиваться конкурентного преимущества. Те, кто раньше и лучше остальных укротит большие данные и внедрит новые аналитические методы, оказывающие значительное влияние на бизнес, будут иметь большое преимущество по сравнению с конкурентами.

Компании, разрабатывающие потребительские товары, имеют специальные лаборатории, которые выводят на рынок новые и инновационные товары. Наличие лабораторий, занимающихся разработкой

инноваций в области аналитики и данных, также очень важно. В сфере анализа больших данных немногие компании имеют опыт, поскольку эта сфера является относительно новой. Однако с точки зрения аналитики, как говорилось в главе 1, большие данные не слишком отличаются от прочих источников данных, которые сначала казались пугающими, но со временем получили широкое распространение. Нет причин, по которым организации не могут начать работу с большими данными уже сегодня.

Создание условий для внедрения аналитических инноваций требует усилий

Как создать в своей организации условия для внедрения инноваций в области аналитики? Это требует согласованных, целенаправленных усилий. Инновациям в области аналитики следует уделять такое же внимание, как и разработке новых продуктов и услуг. Аналитика должна рассматриваться в качестве столпа бизнеса, а не дополнительной функции.

В настоящее время новые источники больших данных появляются столь быстро, что никто за ними не поспевает. В той или иной отрасли очень немногим довелось серьезно поработать с некоторыми из них. Ряд новых источников данных еще только планируется использовать. Дело в том, что большая часть источников больших данных слишком нова и еще не исследована. И все-таки это изменится в ближайшее время. Вопрос в следующем: какие организации станут лидерами, а какие — последователями? И, что еще более важно, где окажется ваша организация?

Традиционные подходы препятствуют внедрению инноваций

В современных организациях, особенно крупных, положение дел определяет бюрократия, которая придерживается давно укоренившихся правил и политики. Процесс внедрения какого-либо нововведения обычно отнимает много времени и сил. Это может быть особенно верно в отношении инновационных способов использования данных и аналитики, поскольку многие люди далеки от этих вопросов в целом. Все это препятствует внедрению инноваций.

Типичный корпоративный процесс характеризуется длительными задержками начала исследования новой идеи. Еще больше времени

уходит на одобрение проекта и его реализацию. Это продолжительный, не терпящий риска, формальный процесс, требующий большого объема документации. От человека, который решил заняться работой с источником больших данных, могут потребовать описания конкретной проблемы, стоящей перед организацией, план проекта, финансовые прогнозы, план по обеспечению проекта персоналом, план проведения анализа, резервный план, оценку рисков и многое другое.

Но вот загвоздка: если вы хотите сделать что-то новое, то у вас по определению не может быть достаточно информации, чтобы с уверенностью подготовить все нужные документы и показатели. Если бы новый способ использования данных для создания дополнительной ценности был полностью понятным, то идея не была бы инновационной. Получить одобрение на реализацию инновационного аналитического проекта может быть очень сложно без разработки отличного от стандартных процессов механизма.

Основное препятствие заключается в невозможности убедить всех в том, что нет риска. С новой идеей обязательно связаны определенные риски. В том-то и дело! Инновационный аналитический проект не свободен от риска. При использовании стандартного процесса идея может быть отклонена. Со временем люди, которые предлагают новые идеи и постоянно сталкиваются с отказами, либо покинут компанию, либо откажутся от новых попыток. Без альтернативы стандартному процессу, который предполагает избегание риска, организации будет трудно внедрять инновации в области аналитики.

Инновации невозможны без риска

Инновационная идея, связанная с анализом больших данных, сопряжена с определенным риском. Очень трудно заставить людей приступить к реализации инициативы, если они не знают, к чему это приведет. Поначалу источник больших данных и способы его использования будут непонятными. Но это не означает, что вам не стоит принимать этот вызов.

Представьте себе, что в типичной корпоративной среде на встрече с руководством аналитическая команда выдвигает новую идею. Аналитики рассказывают о новом источнике больших данных, который никогда раньше не использовался, и говорят, что хотят его исследовать. Они не совсем понимают суть этих данных и то, как компания может их применять, но у них есть много идей, которые они хотели бы опробовать. Тем не менее они уверены в том, что к тому моменту, как они

соберут и проанализируют эти данные, они смогут обеспечить огромную ценность для компании. Эта ценность будет осознаваться в процессе работы. Потребуются некоторые усилия и планирование, чтобы убедить руководителей одобрить этот проект. Существуют ли способы упростить этот процесс? Да. Мы их обсудим.

Понятие аналитической инновации

Определим аналитическую инновацию как новый и уникальный аналитический подход. Это то, чего данная организация не делала прежде, и, возможно, то, чего вообще никто и нигде прежде не делал. Чтобы считаться инновационной, идея должна быть из ряда вон выходящей. Она не может быть вариацией старой идеи. Настройка параметров существующего алгоритма или изменение метода вычисления показателей не инновация. Инновацией может быть, например, использование совершенно иного метода моделирования с применением новых источников данных.

Инновацией не может быть то, что уже предоставляется существующими программными пакетами или сервисами. Возможно, это программное обеспечение или сервисы являются новыми для данной организации, но они не могут считаться инновациями. Другими словами, нечто может являться инновационным подходом для данной организации, но если этот продукт или процесс многими используется, то его применение является не инновацией, а попыткой идти в ногу с рынком.

Внедряйте инновации, а не частичное усовершенствование

Инновационная аналитика подразумевает внедрение чего-то нового и необычного, а не частичное усовершенствование существующего процесса или подхода. Аналитическая инновация должна быть сосредоточена на анализе нового источника данных, решении новой задачи или их комбинации. Конечно, аналитикам следует сконцентрировать свое внимание на областях, которые предоставляют компании больше возможностей.

Инновацией не является проведение разработанного для конкретного случая анализа. Это не значит, что такие вопросы и аналитика не представляют собой ценности. Запрос на проведение анализа и выяснение причин падения объема продаж может обеспечить большую дополнительную ценность. Однако стандартные специальные запросы

предполагают обычную работу, выполняемую аналитической командой. Большая часть специальных запросов не требует использования инновационной аналитики, даже если они ориентированы на решение новой проблемы.

Итеративные подходы к внедрению аналитических инноваций

В целях внедрения аналитических инноваций и исследования больших данных потребуется использовать более подходящие для такой деятельности итеративные подходы. Они делают акцент на сотрудничестве и гибкости. Идея состоит в том, чтобы создать небольшую команду, которая ежедневно работает над обеспечением работоспособности идеи. Команда будет решать возникающие проблемы и в случае необходимости менять направление. Для того чтобы подход способствовал инновациям, он должен учитывать реакцию и быть достаточно гибким: имеет смысл отойти от первоначального плана, если возникает необходимость в изменениях.

Допустим, команда начинает анализировать новый источник больших данных. В процессе работы обязательно возникнут проблемы, связанные с содержимым и структурой данных, в результате чего команде нужно будет скорректировать планы. Основной упор делается не на том, чтобы строго придерживаться изначального плана, а на том, чтобы привести данные в такой вид, в каком они могут быть проанализированы. Возможно, некоторые показатели работают не так хорошо, как предполагалось, но найдены другие, более подходящие. Внесите требуемые изменения. Если в конце первого этапа данные находятся в том виде, который необходим для предстоящего анализа, то это победа независимо от конкретных предпринятых для этого шагов.

Чтобы обеспечить дальнейший успех, после подготовки данных следует все силы направить на то, чтобы быстро создать рабочие прототипы. Цель состоит в том, чтобы получить доказательство работоспособности идеи. Прототип нуждается в достаточном количестве деталей, чтобы люди могли понять общую картину того, что произойдет, если организация возьмется за реализацию идеи всерьез. Возвращаясь к нашему примеру, подчеркнем, что при первом использовании данных не следует стремиться к идеальному анализу. Изначальная аналитика не обязательно должна быть пуленепробиваемой. Скорее, она должна

быть достаточно прочной, чтобы показать плюсы конкретного подхода. Позже будет достаточно времени для совершенствования процесса. Если вы не сможете доказать, что подход имеет свои преимущества, на его дальнейшее развитие не будет выделено времени. А быстро созданный прототип обеспечит такое доказательство.

Ключом к быстрому созданию прототипа являются более короткие итерационные циклы и разбиение задачи на небольшие фрагменты. Это позволяет демонстрировать прогресс; кроме того, решение небольших фрагментов задачи по одному позволяет легче справляться с неожиданностями и учитывать новые сведения, получаемые на каждом этапе.

Проявляйте гибкость

Последствия инновационной аналитической идеи не могут быть полностью осознаны с самого начала. План никогда не будет полностью готов. Команда проекта не будет обладать полным пониманием. Вот почему при исследовании идей очень важно использовать итерационный гибкий подход. Корректировка планов по мере получения новой информации является частью этого праздника.

При исследовании больших данных специалистам в нашем примере потребуется проявлять гибкость. Данные бывают не такими чистыми, как хотелось бы, и не такими полными, как ожидалось; не всегда обладают необходимой прогностической ценностью. Это нормально. Как только аналитики приступят к работе с данными, они могут вдруг осознать, что некоторые следующие запланированные шаги не совсем подходящие. Если они используют гибкий, итеративный подход и не связаны первоначальным планом проекта, они скорректируют планы и двинутся вперед. Пока команда помнит о цели анализа и о том, что нужно доказать, она может гарантировать, что внесенные изменения обеспечивают движение в нужном направлении.

Будьте готовы изменить точку зрения

Пенсионное планирование основано на управлении рисками, которым вы подвергаетесь. В какой-то день в будущем вы планируете выйти на пенсию, и цель пенсионного планирования — обеспечить к этому времени наличие необходимых активов. В основе венчурного инвестирования лежат абсолютно противоположные принципы (см. табл. 10.1). При венчурном инвестировании целью является получение большой

прибыли за относительно короткий период. В то же время эта модель предполагает значительный риск потери всех или большей части вложенных средств.

Табл. 10.1

Пенсионное планирование в сравнении с венчурным капиталом

Пенсионное планирование	Венчурный капитал
Выбор проверенных компаний	Выбор новых компаний
Уравновешивание шансов на получение прибыли и убытков	Акцент на прибыли, а не на убытках
Ориентация на сочетание умеренных результатов	Подход «всё или ничего»
Жестко контролируемая волатильность*	Высокая волатильность
Применяется для пенсионного планирования и управления текущими активами	Применяется для управления активами, не являющимися критически важными для выживания

В данном случае приходится выбирать между риском и доходностью. Оба подхода уместны, если инвесторы понимают, что они делают, и правильно используют доступные возможности. Многие известные компании, особенно работающие в области информационных технологий, электронной коммерции и социальных медиа, финансировались венчурными инвесторами. Венчурный капитал творит чудеса. Однако на каждую компанию вроде Google или Amazon приходится множество компаний, потерпевших неудачу.

Так же как не следует вкладывать все пенсионные накопления в венчурные проекты, нельзя вкладывать все корпоративные ресурсы в рискованные или новаторские инициативы. Необходимо обеспечить выживание компании и такой темп развития, который позволит ей достичь своих целей. Капиталом, выделенным на проведение рискованных операций, нужно управлять. Однако даже пенсионные счета предполагают вложение некоторой части средств в более рискованные активы.

Тем не менее люди редко задумываются об оборотной стороне медали. Большой риск — стремление избегать риска. Например, если 100% средств на пенсионном счету вложены в государственные облигации, то это риск, поскольку доходность таких облигаций вряд ли превысит темп

* Волатильность — статистический показатель, характеризующий тенденцию изменчивости цены.

инфляции. Такой счет не обеспечит достаточной для достижения цели доходности, хотя практически наверняка предотвратит потери. Вполне возможно гарантировать сохранность средств — и одновременно гарантировать то, что у вас хватит денег, чтобы выйти на пенсию вовремя!

Диверсифицируйте!

Организации нуждаются в диверсификации. Насколько разумным является вложение небольшой части пенсионных накоплений в более рискованные активы, настолько же целесообразно выделить небольшой процент корпоративных ресурсов на реализацию аналитических инноваций. Отсутствие инвестиций в инновации может быть столь же рискованным, как и наличие слишком больших инвестиций.

Такому же риску подвергают себя компании, стараясь не рисковать, когда дело доходит до больших данных и аналитики. Если компания хочет в будущем достичь своих целей, ей необходимо принять на себя ответственность за реализацию некоторых рискованных инициатив. Центр аналитических инноваций представляет собой самый подходящий для этого механизм.

Готовы ли вы к созданию центра аналитических инноваций?

Реализации инициатив, связанных с большими данными и углубленной аналитикой в целом, может способствовать центр аналитических инноваций, который предназначен для проведения быстрого исследования идеи и сокращения временного интервала между ее формулированием и реализацией. Центр аналитических инноваций предполагает наличие механизма надзора и фильтрации идей. Это не ведет к анархии, однако значительно ускоряет процессы принятия решения и получения одобрения по сравнению с типичными корпоративными бюрократическими процедурами. Рассмотрим, как все это работает.

Объединение концепций

Центр аналитических инноваций объединяет концепции, изложенные в предыдущих девяти главах. Он подразумевает использование источников больших данных, описанных в первой части; инструментов, процессов и технологий, о которых шла речь во второй части, а также людей и подходов, о которых говорится в третьей части. Создание центра аналитических инноваций является одним из способов, помогающих укротить большие данные и применить к ним новые методы анализа.

Компонент 1: технологическая платформа

Центру аналитических инноваций понадобится технологическая платформа для хранения и анализа необходимых данных, а также отдельная инфраструктура. Команде центра потребуются выделенные ресурсы базы данных, аналитических инструментов, пропускной способности сети и т. д. Для начала следует предусмотреть корпоративное хранилище данных (в идеале), аналитические устройства или витрины данных. Для того чтобы справиться с источниками больших данных, центру нужна система, которая позволит обрабатывать большие потоки полуструктурированных и неструктурированных данных. Как говорилось в главе 4, для этого подходит программный каркас MapReduce.

Эта среда потребует наличия доступа к данным, которые хранятся на существующих платформах. Кроме того, необходимо обеспечить возможность загрузки и анализа данных из новых источников. Многие проекты центра аналитических инноваций потребуют проведения экспериментов с новыми источниками данных в целях проверки целесообразности их объединения с другими данными для создания дополнительной ценности.

Технология центра аналитических инноваций на самом деле довольно проста, учитывая темы, которые мы уже обсудили. С точки зрения реализации инфраструктура центра аналитических инноваций может быть столь же несложной, как аналитическая песочница, о которой шла речь в главе 5. Так же как песочница создается в корпоративном хранилище данных или витрине данных для проведения повседневного анализа, для центра аналитических инноваций может быть создана выделенная среда песочницы. Однако ресурсы, предназначенные для песочницы центра, должны быть отделены от тех, что отведены песочнице для проведения повседневного анализа. Что касается среды MapReduce для центра аналитических инноваций, то для него можно выделить часть существующей среды или создать отдельную среду MapReduce.

Создание инфраструктуры для центра аналитических инноваций не должно вызывать трудностей. Все элементы уже используются в той или иной форме для других видов анализа, производимого организацией. Нужно просто выделить некоторую часть этих ресурсов для центра аналитических инноваций. Учитывая принципы работы песочницы и аналитики, встроенной в базу данных, в покупке совершенно нового набора лицензий или оборудования нет необходимости.

Компонент 2: продукты и услуги сторонних поставщиков

Время от времени центру аналитических инноваций придется использовать продукты сторонних поставщиков, поскольку для эффективной реализации инновационной идеи могут понадобиться дополнительные функции. В имеющихся инструментах иногда отсутствуют необходимые алгоритмы или другие возможности. Например, существует инструмент моделирования, содержащий новый алгоритм, который команда хочет использовать при проведении анализа. В этом случае имеет смысл купить программное обеспечение или пробную лицензию для данного проекта.

Организации может потребоваться помощь внешних консультантов, обладающих опытом работы в исследуемой предметной области. Ни одна организация не содержит штатных специалистов, имеющих все навыки, которые могут пригодиться в процессе работы. Это особенно верно, если организация собирается выйти на новый уровень. Внутренние ресурсы не всегда располагают свободным временем. Шансы на успешную реализацию инновационной аналитической идеи можно значительно повысить при соответствующей поддержке внешних экспертов, которые уже знают, что нужно делать.

Компонент 3: обязательность и спонсорство

Центру аналитических инноваций придется принять обязательства по реализации новых идей с использованием итерационных процессов, для чего потребуются спонсорство и участие руководителей высшего звена. В конце концов, такой центр представляет собой значительные инвестиции, которые некоторые люди сочтут довольно рискованными. Только руководитель может оказать поддержку, необходимую для того, чтобы получить одобрение и справиться с бюрократией.

Не менее важно наличие спонсоров для каждого отдельного проекта центра. Спонсоры проекта должны принимать активное участие в работе над ним, а также иметь полномочия, способности и намерение реализовывать одобренные идеи. Это крайне важно, так как нет смысла доказывать, что идея имеет огромный потенциал, если никто не собирается ее реализовывать. Без поддержки и видения руководителей проект может оказаться технически успешным, но не приведет к положительным изменениям.

Компонент 4: сильная команда

Центру аналитических инноваций нужна команда, обладающая навыками в сфере бизнеса, ИТ и аналитики. Наиболее важной частью центра является выделенная для него команда специалистов. Специалисты в области бизнеса помогут определить проблемы и учесть практические соображения. Аналитики разберутся в данных и выполнят необходимый анализ. ИТ-специалисты будут управлять большими объемами данных, оборудованием и процессами, используемыми центром.

Очень важно, чтобы в центре аналитических инноваций работали самые лучшие профессионалы. Работа в центре должна почитаться за честь. Со временем люди могут приходить и уходить, но в идеале часть сотрудников будет работать в центре на постоянной основе. Если это невозможно из-за ограниченного бюджета, то работа в центре должна стать формальной частью должностных обязанностей членов команды. Если людям необходимо тратить 20% своего рабочего времени на работу в центре аналитических инноваций, то эти 20% времени следует официально выделить на данную работу. Если же сначала команда должна выполнить все свои обязанности, а при наличии свободного времени — работать над проектами центра, то это не сработает. Ничто никогда не сдвинется с места, потому что для этого никогда не будет времени.

Компонент 5: совет по инновациям

Последним компонентом центра аналитических инноваций является совет по инновациям. В него должны входить члены команды центра, его спонсоры из числа руководителей, а также представители соответствующих бизнес-единиц. Совету предстоит рассматривать идеи и выбирать те, которые следует реализовать. Подобно тому как идея нового бизнеса представляется венчурным инвесторам, кто-то может сделать краткую презентацию новой идеи, ее ценности, а также плана ее исследования и создания прототипа. Совет слушает, задает вопросы, а затем обсуждает представленную идею. Затем определяет целесообразность ее реализации. Если идея имеет ценность, то ее добавляют в список проектов.

Совету предстоит наблюдать за реализацией проектов. На обсуждение совета выносятся промежуточные результаты, а также возникшие проблемы. Совет предоставляет идеи относительно реализации

проекта по мере появления новой информации. Если проект сталкивается с серьезным препятствием, совет может принять решение об отказе от дальнейшей работы. Далее мы поговорим о том, как справляться с неудачами. Суть в том, что совет находится в курсе всего, что происходит в центре, и направляет его деятельность к реализации наиболее ценных идей.

Руководящие принципы центра аналитических инноваций

Итак, центр аналитических инноваций должен быть автономным и гибким и обладать ресурсами, необходимыми для достижения успеха. Выделенную среду не следует ограничивать производственными потребностями или процессами — в этом поможет физическое или логическое разделение. Это не означает, что центр может использовать все доступные системные ресурсы. Смысл в том, что если решено выделить 10% доступных ресурсов, то эти 10% должны быть доступны для работы процессов центра. Сотрудникам центра необходимо выделить время на разработку и реализацию идей центра. Работа над проектами центра аналитических инноваций должна являться не дополнительной нагрузкой, а частью повседневных рабочих обязанностей.

При определении основных правил работы центра, включая работу совета по инновациям, необходимо фокусироваться на быстром реагировании и минимизации бюрократии и волокиты. Если в центре работают высококлассные специалисты, нет смысла тормозить их деятельность. Волокита призвана останавливать нечестных или некомпетентных людей. Честные, высококвалифицированные специалисты не нуждаются в волоките, поскольку они все делают правильно. Разумеется, это относится и к профессионалам в области аналитики, о которых мы говорили в главе 8.

Важным руководящим принципом является то, что центр аналитических инноваций должен следовать модели венчурных инвестиций, а не пенсионных накоплений. Для реализации следует выбирать проекты, характеризующиеся потенциалом большой положительной доходности и продвигающие одну или несколько стратегических инициатив компании. Некоторые идеи потерпят поражение, это нормально. Цель в том, чтобы осознать это поражение как можно быстрее. Необходимо выявить несколько удачных идей из множества неудачных и не переживать, если что-то не получится.

Сфера деятельности центра аналитических инноваций

Центр аналитических инноваций не должен заниматься решением заурядных вопросов. Если вы заметили, что ваш центр занимается обычным анализом данных или созданием специализированных отчетов, остановите это! Если ваш центр тратит время на незначительное усовершенствование продуктов, остановите это! Если ваш центр тестирует новое приложение просто для того, чтобы посмотреть, как оно работает, остановите это! Если ваш центр работает над формальной реализацией аналитического процесса, остановите это! Даже в случае с очень успешным проектом центру аналитических инноваций потребуется время, чтобы доказать, что изучаемая концепция может и будет работать. Этот центр не должен быть частью долгосрочного плана.

Используйте центр аналитических инноваций по назначению

Центр аналитических инноваций должен использоваться только для тех целей, для которых он предназначен. Первоначальные исследования, сфокусированные исследования и начальные прототипы — вот его основные задачи. Планирование более масштабной и формальной реализации прототипа выходит за пределы сферы деятельности центра. В этот момент ресурсы центра должны быть направлены на решение следующей проблемы, а проект необходимо передать команде, которая будет заниматься его реализацией.

Процесс внедрения аналитических инноваций подразумевает несколько этапов. На рис. 10.1 ясно видно, где заканчивается работа центра аналитических инноваций. Процесс начинается с появления идеи. Сначала производятся некоторые начальные исследования для ознакомления команды с проблемой. Далее происходит более глубокое, сфокусированное исследование для того, чтобы добраться до сути проблемы. Это приводит к разработке прототипа, который показывает, как идея может работать. Вплоть до этого момента участие центра аналитических инноваций является целесообразным. Как только приходит время для создания формальной версии прототипа, предназначенного для развертывания, центр должен прекратить работу над проектом. После получения доказательства работоспособности идеи ее реализацию следует направить по более традиционному пути, но в ускоренном режиме.

Центр аналитических инноваций задействуется только на начальных этапах разработки идеи. На каждом этапе происходит дальнейший отбор проектов

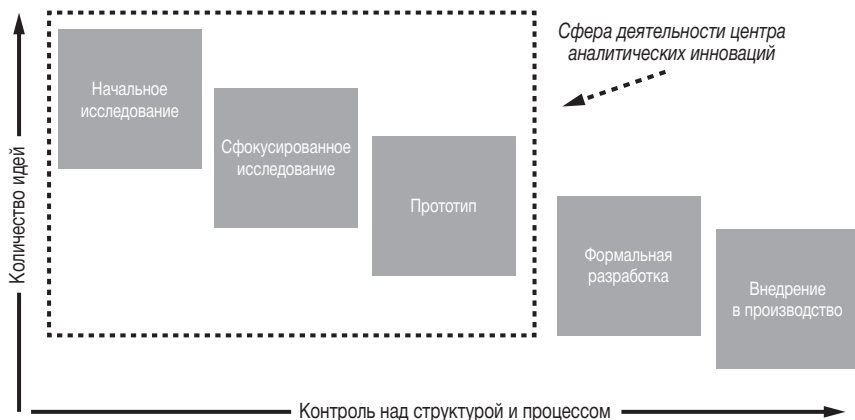


Рис. 10.1. Сфера деятельности центра аналитических инноваций

Существует вероятность того, что после успешного создания прототипа проект столкнется именно с теми бюрократическими препятствиями, во избежание которых и создается центр. Тем не менее, когда приходит время принимать дорогостоящие решения, следует ограничить риски и убедиться в том, что данная идея реализуется в соответствии с другими корпоративными инициативами. Центр аналитических инноваций позволяет сократить путь от формулирования идеи до подготовки к реализации. Ключом к успеху является дальнейшая ускоренная разработка идеи, ограниченная только необходимым тестированием и изменением процессов.

Рассмотрим для примера организацию, которая решила исследовать обращения в центр обслуживания клиентов, чтобы определить настроение и выявить слабые места продукта. Сначала команде центра аналитических инноваций дается выборка электронных писем. С помощью пробных версий инструментов для анализа текста проводится предварительное исследование. На данном этапе команда находит некоторые убедительные результаты, что обеспечивает поддержку реализации идеи в широком масштабе.

Реализация идеи выходит за рамки сферы деятельности центра и должна являться формальным проектом. Необходимо приобрести лицензии на программное обеспечение для анализа текста, обеспечить

постоянный поток электронных писем, а также усовершенствовать аналитические методы, используемые в прототипе. Эти этапы не должны являться частью работы центра инноваций, поскольку его задача считается выполненной, как только будет доказано, что идея анализа электронных писем клиентов обладает достаточной ценностью. Предполагается, что доказательства, предоставляемые центром аналитических инноваций, способствуют быстрой реализации идеи. Поскольку ценность идеи доказана, на ее внедрение должны быть выделены средства.

Следует упомянуть об одном исключении, касающемся ограничений сферы деятельности центра. Рассмотрим случай, когда центр аналитических инноваций создал новый мощный аналитический процесс. Он доказал, что новый источник больших данных может значительно помочь бизнесу. Центр даже знает, что необходимо сделать для введения этого процесса в эксплуатацию, и уже разработал соответствующий план. В этом случае целесообразно пересмотреть границы сферы деятельности центра аналитических инноваций.

Бессмысленно сообщать бизнес-сообществу о новом потрясающем аналитическом процессе, а также о существовании в центре аналитических инноваций прототипа, если во время его ввода в эксплуатацию процесс не будет работать, а запустится через три-шесть месяцев и только тогда пользователи начнут получать результаты. В данном случае можно обеспечить работу процесса в центре аналитических инноваций в период его формальной реализации, чтобы он мог приносить пользу сразу же.

Применение описанного исключения требует дисциплины. Если организация не намерена быстро ввести процесс в эксплуатацию, то она рискует загрузить центр аналитических инноваций работой псевдопроизводственного процесса. Это неправильно. В частности, если процесс работает в центре аналитических инноваций, то его реализация может откладываться. Время может быть потрачено на другие проекты, поскольку по крайней мере спонсоры центра получают свои результаты в данный момент. Центр аналитических инноваций как можно скорее должен освобождаться от идей, ценность которых им уже доказана, и переходить к следующей идее. В то же время не стоит лишать бизнес-спонсоров необходимой им информации, буквально следуя правилам и прекращая работу после создания начального прототипа.

Как быть с неудачами

Центр аналитических инноваций будут время от времени постигать неудачи. Не каждая идея может сработать. Неудачи возможны на разных уровнях.

К первому типу неудач относится полный провал. Это, например, происходит, когда все сочтут, что команда ошиблась изначально, решив, что идея может сработать. Или идея была отличная, но в данных условиях она просто не сработала. Неудачи такого рода приводят к отказу от идеи. Однако даже в таком случае о данных можно узнать много нового и все эти новые сведения могут быть использованы в будущем. То, что идея, выдвинутая центром аналитических инноваций, не сработала, еще не означает, что в процессе ее исследования не была получена ценная информация.

Ко второму типу неудач относятся случаи, когда идея не срабатывает в настоящее время, но команда считает, что ее удастся реализовать в дальнейшем, когда будут доступны новые данные или более мощная система, которая сможет справиться с обработкой, или корректировка бизнес-модели компании позволит внести в процесс необходимые изменения. Работа над идеей приостанавливается до устранения существующих препятствий. Как и в случае с первым типом неудач, полученные уроки могут быть незамедлительно применены к другим проектам.

К последнему типу неудач относятся случаи, когда идея на самом деле срабатывает. Она обладает определенными преимуществами и может принести компании прибыль, однако возврат недостаточно высок, чтобы оправдать требуемые усилия, или существуют другие проекты, которые могут обеспечить более высокий возврат. Реализация данной идеи либо откладывается, либо отменяется.

В основу работы с неудачами положены три принципа. Во-первых, целью любого проекта является принятие минимального риска, необходимого для того, чтобы оценить потенциал идеи. Суть центра аналитических инноваций будет утеряна, если ему поручить работу над многомиллионными проектами. Работа над проектами центра должна занимать от нескольких недель до нескольких месяцев. Если для того, чтобы полностью решить проблему, необходимы достаточно серьезные усилия, поручите центру начальный, более короткий этап, который позволит определить целесообразность продолжения работы.

Во-вторых, не беритесь за проекты, которые настолько глобальны, что компания не сможет справиться с чередой неудач, если центр попадет в полосу невезения. Этот принцип можно проиллюстрировать на примере профессиональных игроков в покер.

Одним из золотых правил в профессиональном покере является то, что играть следует только тогда, когда при данном уровне ставок вы можете позволить себе потерять все, что ставите. Ни один игрок в покер никогда не ставит все свои деньги по причине вероятности выпадения таких карт, которые приведут к потере всех или большей части денег в одной игре.

Однажды я был свидетелем того, как четыре туза проиграли комбинации из пяти карт одной масти, идущих по порядку. Шансы реализации такого сценария крайне малы. Большинство людей потеряет все свои деньги, когда у них в руках четыре туза, поскольку в данном случае они готовы поставить все, что у них есть. В конце концов, когда вы держите в руке четырех тузов, а остальные игроки уравнивают вашу ставку, то вы считаете, что они просто отдают вам свои деньги. С точки зрения статистики это имеет смысл, поскольку шансы потери бесконечно малы. Тем не менее потеря возможна.

Дело в том, что профессиональные игроки в покер принимают такой риск, который позволяет им пережить полосу неудач, после чего статистические вероятности гарантируют, что они в конечном счете попадут в полосу везения и отыграются. Худшее, что можно сделать, — это проиграть все непосредственно перед окончанием череды неудач.

Если у вас не бывает неудач, то вы не преуспеваете!

Если центр аналитических инноваций никогда не испытывал неудач и если некоторые из этих неудач не являлись полным провалом, то это означает, что центр слишком старательно избегает риска. Инновации невозможны без риска. Понять, почему один аналитический процесс не сработал, может быть столь же полезно, как понять то, что же именно обеспечило успех другого процесса.

Третий и, вероятно, самый важный принцип заключается в том, чтобы помнить: неудача многому может научить. Неудачу следует воспринимать не как провал, а как опыт. Теперь вы знаете, почему первоначально предполагалось, что идея работает, а также то, почему этого не произошло, и это может оказаться чрезвычайно полезным при разработке процессов в будущем. Если неудачный проект позволил выявить новое ограничение, относящееся к данным, или некорректную поста-

новку проблемы, можно принять меры, чтобы это не мешало функционированию других процессов. Новые знания могут повлиять как на текущую, так и на будущую работу.

По мере того как организация реализует концепцию центра аналитических инноваций и осознает его ценность, он может быть расширен за счет выделения дополнительных ресурсов, людей и технологий. Приобретенный опыт может быть применен к другим областям бизнеса. Центр инноваций нужен не только для сферы аналитики, но и для других сфер. Как только организация наладит инновационную деятельность в одной области, ей будет проще перенести этот опыт на другие.

Обзор главы

Самые важные уроки этой главы.

- ▶ Инновации в сфере аналитики и использования данных представляют собой важную область деятельности организации.
- ▶ Инновации в сфере аналитики должны быть сосредоточены на анализе новых источников данных, решении новых проблем или их комбинации. Инновации не являются простым расширением существующей деятельности.
- ▶ Инновационная идея подразумевает риск и не является вполне понятной. Аналитические инновации требуют применения итерационного, гибкого подхода. Планы необходимо корректировать по ходу дела.
- ▶ Некоторая часть корпоративных ресурсов должна использоваться в соответствии с моделью венчурного капитала, а не с моделью пенсионных накоплений. Центр аналитических инноваций — самый подходящий для этого способ.
- ▶ Центр аналитических инноваций подразумевает наличие людей, процессов и технологий, а также кросс-функционального наблюдательного совета, направляющего деятельность центра по правильному пути.
- ▶ Очень важно, чтобы у людей, работающих над проектами центра, на это была официально выделена часть рабочего времени. Работа в центре не может быть дополнительной нагрузкой.

- Не загружайте центр аналитических инноваций производственными процессами или созданием окончательной рабочей версии разработки. Сфера деятельности центра аналитических инноваций ограничивается созданием прототипа.
- Рискуйте только той частью ресурсов, которая позволяет оценить потенциал идеи, и не беритесь за крупные проекты, если череда неудач при их реализации может оказаться разрушительной.
- Неудачи в центре аналитических инноваций необходимо выявлять как можно быстрее, чтобы команда могла перейти к решению других проблем.
- На неудаче можно многому научиться. Она не является однозначно плохим явлением. Неудачи весьма полезны, если полученные знания широко применяются для улучшения других процессов.

Создание культуры инноваций и открытий

Если у вас есть дети, можете ли вы сказать, какую игрушку они больше всего захотят получить на свой следующий день рождения? Если у вас нет детей, можете ли вы сказать, за каким новым модным электронным устройством люди будут стоять в очереди, как это было в случае с iPhone™ или Wii™? Можете ли вы сказать, какой будет следующее совершенно нелепое увлечение, вроде Pet Rocks*?

Большинство людей скажут, что не имеют об этом ни малейшего представления. И это нормально, поскольку инновационные подходы не бывают очевидными, пока вы не увидите их. Если бы идеи были очевидными, они не были бы инновационными. Если бы кому-то регулярно удавалось предвидеть такие нововведения, то он уже удалился бы на свой красивый остров, чтобы наслаждаться честно заработанным богатством!

В этой главе мы подведем итоги книги и поговорим о создании культуры инноваций и открытий. Эта глава будет веселой и беззаботной, но также предоставит и пищу для размышлений. Большинство концепций обсуждают довольно часто, но их стоит пересмотреть, учитывая, как организация может применить эти общепринятые принципы к сфере больших данных и углубленной аналитики. Кстати, если вы любите Pet Rocks, ознакомьтесь с современным вариантом USB Pet Rocks на сайте ThinkGeek.com²⁰¹

* Pet Rocks — «каменные питомцы» — разрисованные в виде животных камешки. Это увлечение появилось в 1970-х годах, позже их аналогом стали тамагочи.

Подготовка почвы

Silly Bandz® и Jibbitz™ — только два свежих примера инновационных продуктов, которые появились из ниоткуда и завоевали рынок. На случай, если вы не знаете, Silly Bandz — это разноцветные резинки в форме животных, игрушек или других предметов. Дети обожают их носить, обмениваться ими и коллекционировать их. Они настолько популярны, что были запрещены во многих школах, поскольку отвлекают внимание от занятий. Jibbitz — это маленькие фигурки, которые вставляются в отверстия в обуви Crocs™. Посмотрите на кроксы*, которые носят дети, и вы увидите, что в большинстве случаев на них будет несколько фигурок Jibbitz. Silly Bandz и Jibbitz — примеры инновационных идей, завоевавших огромный успех.

Подобные инновации не являются чем-то абсолютно новым. Тысячелетия назад Пифагор выдвинул идею о том, что Земля круглая, а не плоская. Много веков назад Коперник заявил, что центром Солнечной системы является Солнце, а не Земля. Несколько десятилетий назад Эйнштейн вывел общую теорию относительности.

Одни нововведения вполне самостоятельны, другие появляются как результат развития предыдущих инноваций. Например, Jibbitz не имели бы никакого смысла, если бы не существовало обуви Crocs. В каждом таком случае кто-то должен проявить смелость, чтобы озвучить новую сумасшедшую идею, а затем реализовать ее. Для успешной инновационной деятельности организации предстоит развить культуру инноваций и открытий, которая поощряет эксперименты, побуждает людей бросать вызов предположениям и способствует изменению правил ведения дел. Это не просто, это требует усилий и сосредоточенности. Однако без такой культуры организация будет медленно угасать по мере того, как конкуренты будут неумолимо отнимать ее бизнес.

Осознала ли ваша компания ценность больших данных и прилагает ли усилия для их укрощения? Разработала ли ваша компания новые аналитические процессы, подразумевающие использование больших данных? Есть ли у вашей компании видение того, как большие данные можно объединить с другими данными, чтобы создать дополнительную ценность? Опробовала ли ваша компания некоторые из новых

* Crocs — всемирно известный бренд взрослой и детской обуви.

«сумасшедших» аналитических идей за последнее время? Вне культуры инноваций и открытий шансы получить утвердительный ответ сильно уменьшаются.

Невероятная история Crocs и Jibbitz

Только представьте себе разговор, который состоялся в процессе изобретения обуви Crocs. Однажды кто-то говорит: «У меня есть отличная идея! Давай сделаем твердые резиновые башмаки. Они не будут красивыми и дешевыми, однако люди с радостью будут носить их в бассейнах, во время работы в саду и т.д. А детям они понравятся потому, что им не попадет, если они их испачкают. Родители просто польют их из шланга». Многие компании отбросили бы эту идею и не рискнули ее реализовывать. Большинство людей не поверили бы в успех, которого достигла обувь Crocs.

А у изобретателя Jibbitz есть пара обуви Crocs, и он понимает: «В этой обуви есть небольшие отверстия. Надо их чем-то закрыть! Я сделаю для этих отверстий маленькие клипсы с фигурками персонажей мультфильмов, флагами или другими предметами. Наверняка детям понравится украшать ими свою обувь. Это поможет им найти свою пару среди четырех пар красных кроксов на детской площадке. Лучше всего то, что их производство будет стоить очень дешево, но мы будем продавать их по цене \$3, \$4, может быть, даже \$5 за штуку. Мы сорвем большой куш!»

Вы бы бросили основную работу, чтобы рискнуть своими сбережениями ради реализации этой идеи? В конце концов Jibbitz завоевали ошеломительный успех. Рич и Шери Шмельцер основали компанию по продаже Jibbitz в 2005 году. Спустя год они продали ее компании, которая производит обувь Crocs, за 20 миллионов долларов. Вот это успешная инновация!

Вернемся в мир аналитики и рассмотрим идею программного пакета, предназначенного для статистики и анализа. Такой вещи не существовало 40 лет назад. В 1976 году после того, как программное обеспечение SAS добилось некоторого успеха в качестве проекта, финансируемого Университетом Северной Каролины, несколько профессоров оставили свою работу, чтобы создать новую компанию SAS Institute. Эта компания должна была заниматься «поддержкой и дальнейшим развитием SAS»²¹. Кто мог предположить, что она превратится в многомиллиардную организацию, которая будет играть важную роль в бесчисленных аналитических инновациях?

Эффективная инновационная деятельность

Задумайтесь на минуту о том, какие факторы должны были сойтись во-едино для того, чтобы Jibbitz добились успеха. Шмельцерам предстояло разработать совершенно новую идею, связанную с заполнением отверстий в обуви Crocs. Затем они должны были поверить в ее потенциал и полностью посвятить ее реализации свое время. Путь от создания концепции Jibbitz до их производства и продажи потребовал серьезных усилий. Двадцать миллионов долларов, безусловно, не заставили их пожалеть о том пути, который они выбрали.

Может ли каждая идея оказаться столь же успешной, как Jibbitz? Конечно же, нет. В этом суть инноваций. Вы не можете выигрывать каждый раз. Это нормально. Проблема в том, что вы никогда не достигнете такого успеха, как Jibbitz, если не будете пробовать реализовывать новые идеи. У вас нет гарантий выигрыша, если вы будете играть, но, чтобы выиграть, вам обязательно надо играть. Шмельцеры сыграли и выиграли. Сделали бы вы и ваша организация то же самое?

Чтобы выиграть, нужно играть

Не все нововведения окажутся успешными. Выигрыш в игре в инновации не гарантирован. Однако если вы не играете, то точно не выиграете. Чтобы выиграть, нужно играть! Если вы не прилагаете усилия к тому, чтобы укротить большие данные, то ваша организация не сможет их укротить, а также воспользоваться теми преимуществами, которые связаны с их анализом.

Что вы и ваша компания сделали в последнее время, чтобы повысить шансы на успех в области углубленной аналитики и больших данных? Сколько энергии люди вкладывают в обдумывание новых аналитических процессов в день, в неделю и в месяц? Кто в вашей компании отвечает за создание культуры инноваций и открытий? Если неясно, кто занимается продвижением аналитических инноваций в организации, то, очевидно, этому не уделяется должного внимания. Если никто не несет за это ответственности, то существует вероятность, что никто не будет ничего делать. Хотите ли вы быть человеком, отвечающим за инновации в области аналитики в вашей организации? Поговорим об этом.

Обзор ключевых принципов

Существует три ключевых принципа, которым необходимо следовать для того, чтобы создать культуру инноваций и открытий, способству-

ющую внедрению аналитических инноваций и укрощению больших данных:

- ▶ **Принцип 1: выйдите за рамки.** Это означает, что революционные инновации — это не простое усовершенствование старой концепции. Они ломают стереотипы и заставляют делать дела по-новому.
- ▶ **Принцип 2: воспользуйтесь волновым эффектом.** Это означает, что лучшие инновации, будучи усовершенствованными, позволяют осуществлять другие нововведения, даже если они не были запланированы. Одна инновация порождает другую, та порождает третью, а совокупный эффект может затмить эффект первоначальной инновации.
- ▶ **Принцип 3: сосредоточьте все силы на достижении цели.** Руководители организации должны определить общее видение и установить четкие приоритеты. Они должны связать зарплаты сотрудников с этим видением и приоритетами. Если все силы не будут сосредоточены на одной цели, то создать культуру инноваций и открытий невозможно.

Хотя эти принципы и не новы, они не потеряли своего значения. Время от времени следует возвращаться к основам. Давайте глубже разберемся в каждом из этих широко применяемых принципов и в том, как они используются в мире больших данных и углубленной аналитики!

Принцип 1: выйдите за рамки

Первый ключевой принцип заключается в том, чтобы выйти за рамки. Каждый из нас ограничен какими-то рамками. Это верно как для личности, так и для команды и компании. Вы можете забывать об этом, но эти рамки постоянно вас ограничивают. Они определяются тем, что вы можете и чего не можете делать; тем, что вам следует и не следует делать; тем, что, по вашему мнению, сработает и не сработает, а также тем, что получит и не получит одобрения. Ваши рамки определяются бюджетами, опытом, традициями, навыками, системами, доступностью данных и многими другими факторами.

В какой-то степени эти рамки являются положительным явлением. Они помогают нам сохранить сосредоточенность и добиться своей

цели. Они помогают учитывать важные практические соображения в процессе повседневной работы. Однако если эти рамки жесткие, а не гибкие, то вы сталкиваетесь с проблемами. Если вы регулярно не пересматриваете свои рамки, они становятся серьезной помехой. Смысл состоит в том, чтобы убедиться, что вы искусственно не ограничиваете себя и свою организацию.

Что изменилось с тех пор, когда вы в последний раз определяли, что может и не может быть сделано, что следует и не следует делать, а также что принято и не принято в вашей организации? Когда в последний раз думали о том, почему ваша организация использует именно такую аналитику? Когда вы в последний раз пересматривали принципы анализа, о которых шла речь в главе 7, чтобы убедиться, что ваша аналитическая команда находится на верном пути? Когда в последний раз рассматривали возможность обновления процессов и подходов, о чем говорилось в главах 5 и 6? Когда в последний раз выполняли обновление системы, чтобы воспользоваться преимуществами текущих уровней масштабируемости, описанных в главе 4?

Когда вы в последний раз пересматривали рамки?

Действовать в определенных рамках не всегда плохо. Однако вы должны постоянно следить за тем, актуальны ли существовавшие ранее ограничения. В противном случае вы будете ограничивать себя без необходимости. Раздвигайте реальные, а не предполагаемые пределы возможного. Возможность масштабирования аналитических процессов улучшилась в последние годы так сильно, что многие организации даже не осознают тех потенциальных преимуществ, которые они упускают.

Недавней тенденцией в аналитике является феномен публичных конкурсов аналитиков. В этих конкурсах набор данных, очищенный от любой чувствительной информации, предоставляется тому, кто его попросит. Определяются правила, касающиеся нужного типа модели или анализа, а также критерии выбора победителей. Команда, которая достигла наилучших результатов, по итогам конкурса получает денежный приз. Давая экспертам всего мира возможность применить свои методы, организации могут полностью выйти за свои рамки. Люди со стороны не имеют тех предвзятых суждений, которыми обладает сама организация. В итоге участники конкурсов часто показывают результаты, превышающие исходные результаты проводящей конкурс организации.

Если вы хотите вывести свою организацию за рамки, нужно, чтобы ваши сотрудники бросали друг другу вызов, постоянно пересматривали предположения и убеждались, что они раздвигают реальные, а не предполагаемые границы возможного.

Берите пример с Коперника

Коперник — замечательный пример выхода за рамки. В XV–XVI веках Земля считалась центром Солнечной системы. Коперник проанализировал все доступные данные и решил, что в центре Солнечной системы находится Солнце, а не Земля. Его инновационная идея открыла новую эру в астрономии. Считается, что Коперник положил начало научной революции, которая привела к отказу от доктрин, преобладавших со времен Древней Греции²². Научная революция заложила основы современной науки.

Тем не менее после того, как Коперник произвел свои изыскания, он не торопился представить их на суд общественности. Он понимал, какие споры вызовет такая инновационная идея. Старая модель предполагала, что центром Солнечной системы была Земля, и это предположение было ошибочным. Его новая модель не могла получить признание тех людей, которые были заинтересованы в старой модели.

Если вы используете неправильную модель, то результаты будут частично или полностью лишены смысла. Пример Коперника показывает, что даже успешные инновации не всегда признаются сразу. Копернику потребовалось несколько лет, чтобы опубликовать результаты своих исследований. Его теории стали общепринятыми почти через три века. В конце концов истинность его инновационной идеи восторжествовала. Имейте в виду, что Коперник признавал Землю важной частью Солнечной системы. Она просто не была ее центром. Заявление о том, что Земля не является центром, еще не означает, что она не играет важной роли.

Сосредоточьтесь на людях, а не на инструментах

Коперник показал, что центром Солнечной системы является Солнце, а не Земля. Сегодня многие организации сосредотачивают внимание на технологиях и инструментах, считая их самыми главными факторами достижения успеха в области аналитики. На самом деле главным фактором являются не инструменты и технологии, а люди, которые их используют. Пока вы не сосредоточитесь на правильной модели, ваш успех будет значительно ограничен.

Одной из распространенных ошибок в мире бизнеса является то, что организации считают такие инструменты, как базы данных и программное обеспечение, центром экосистемы углубленной аналитики. Многие организации заявляют, что у них самые лучшие инструменты, программное обеспечение или базы данных. Все это, безусловно, очень важно. Организация, использующая самые лучшие инструменты, программное обеспечение и системы, имеет преимущества. Инструменты и технологии — чрезвычайно важные факторы, определяющие способность любой организации внедрять аналитические инновации.

Однако что заставляет эти технологии стимулировать рост бизнеса? Люди, применяющие их; способ, который они при этом используют; а также решения, принятию которых они способствуют. Люди, использующие инструменты, программное обеспечение и системы, находятся в центре аналитической экосистемы — об этом мы говорили в главах 8 и 9. Организации необходимо удостовериться в том, что она предоставляет нужные инструменты нужным людям для получения нужных результатов. Даже самая мощная система с самыми сложными инструментами не в состоянии укротить большие данные, если процессом не будут управлять выдающиеся профессионалы в области аналитики.

Применение принципа

Если современная организация стремится к инновациям, то аналитика должна стать главной частью фундамента. Если организация поставила цель выйти за рамки и укротить волну больших данных, то ей необходимо использовать последние инновации в области аналитики. Одной из этих инноваций является обработка информации в базе данных для обеспечения масштабируемой аналитики, о чем мы рассказали в главе 4. Как это часто бывает, тем, кто продвигал эту идею еще в 1990-х годах, пришлось многое вынести. Большинство людей, в том числе работающих в сфере аналитики, первоначально встретили данную концепцию с подозрением и презрением.

Важно понимать, что аналитика, встроенная в базу данных, в настоящее время становится общепринятым инструментом. Компании, которые работают в различных отраслях по всему миру, почувствовали значительное увеличение скорости, которое обеспечивает данная технология. Мы говорим об ускорении работы аналитических процессов в 40, 50 и даже в 100 раз по сравнению с применением традиционных подходов. Это не просто частичное усовершенствование —

это скорее квантовый скачок. Безумием было бы не воспользоваться такой возможностью сегодня! Вы легко можете найти в интернете примеры, начиная с выступлений на конференциях и статей и заканчивая формальными описаниями бизнес-ситуаций, которые показывают, как с помощью аналитики, встроенной в базу данных, компаниям удалось вывести аналитическую масштабируемость на новый уровень. Такая же история развивается и вокруг модели MapReduce, о которой шла речь в главе 4.

Представьте себе, сколько дополнительного времени команда аналитиков сможет посвятить инновациям, если будет применять аналитику, встроенную в базу данных, и модель MapReduce для ускорения работы процессов. Многие организации даже не догадываются, что такие результаты можно получить уже сегодня. Они все делают по-старому с теми же ограничениями, а тем временем конкуренты их опережают. Рамки возможностей с точки зрения аналитической масштабируемости раздвинулись. Убедитесь в том, что ваша организация использует дополнительное пространство.

Принцип 2: воспользуйтесь волновым эффектом

Если вы не выходите за рамки и не стремитесь к инновациям, то вам недоступен волновой эффект. Многие инновации со временем оказываются еще более ценными, так как способствуют другим инновациям. Второй принцип заключается в использовании волнового эффекта. Он равнозначен возведению инноваций в квадрат!

Возведите инновации в квадрат!

Инновации способны повлечь за собой непредвиденные нововведения в будущем, которые усиливают эффект исходной инновации. Большие данные настолько новы, что их будущий волновой эффект все еще далеко не ясен. Однако он, безусловно, проявит себя. Не упустите шанс воспользоваться им и приступайте к анализу больших данных уже сегодня!

По-настоящему инновационные идеи часто приводят к другим непредвиденным прорывам, которые имеют даже большее значение, чем исходная инновация. Потенциальные последствия инноваций нельзя сбрасывать со счетов. Не внедряя инновации сегодня, вы упускаете не только непосредственные, но и потенциальные возможности, которые могли бы привести к значительным изменениям в будущем.

Слишком скорый отказ от идей и внедрения инноваций может в долгосрочной перспективе обойтись намного дороже, чем можно себе представить. Рассмотрим несколько примеров.

От телефонов к интернету и социальным медиа

Телефон полностью изменил способ общения людей и оказал огромное влияние на общество. Телефонные линии были протянуты по всему миру с единственной целью — позволить двум людям поднять трубки и поговорить. В то время не было более масштабного видения. Если бы кроме обеспечения возможности общения из этих усилий ничего бы не возникло, то телефон по-прежнему считался бы одним из наиболее значительных изобретений в истории.

Однако кто-то пришел к мысли, что по этим же линиям можно передавать данные. Сначала появились простые факсы, а со временем стал широко использоваться модем. Те же самые старомодные телефонные линии позволили получить доступ к ранней версии интернета. Интернет в том виде, в котором он существует, не мог бы возникнуть в принципе или возникнуть так быстро, если бы критическая масса людей не имела к нему доступа. Эта критическая масса создала спрос, который вызвал взрывное развитие интернета. А началось все с простых низкотехнологичных телефонных линий! Интернет — это пример волнового эффекта, обеспеченного распространением телефонов. Когда первые линии были введены в действие, никто и не предполагал, к чему это приведет, однако они используются совершенно иным способом. Вот что такое волновой эффект.

От телефона к социальным медиа и аналитике

Телефон произвел революцию в сфере коммуникаций. Телефонные линии позднее сыграли огромную роль в развитии интернета. А интернет породил социальные медиа, которые совершили революцию в том, как мы общаемся. Представьте, какой была бы жизнь сегодня, если бы не эти телефонные линии!

Волновым эффектом телефона стал интернет, а он породил собственный волновой эффект. Даже в середине 1990-х годов мало кто мог себе представить все нововведения, которые появятся благодаря интернету. Волновым эффектом является взрывное развитие электронной коммерции, включая бизнес-модели таких прорывных компаний, как Amazon,

eBay и Craigslist. В качестве недавнего примера волнового эффекта можно назвать такие социальные сети, как LinkedIn и Facebook. Они предлагают новые способы общения. Как это ни парадоксально, телефоны произвели революцию в том, как мы общаемся, затем телефонные линии способствовали развитию интернета, а интернет, в свою очередь, революционно изменил способ нашего общения, породив социальные медиа.

К волновым эффектам интернета относятся массовые многопользовательские игры, которые позволяют игрокам в режиме реального времени взаимодействовать с другими людьми в разных уголках планеты. Еще одним эффектом является мобильный телефон с функцией GPS, который дает возможность найти все китайские рестораны на расстоянии километра от вашего текущего местоположения и предусматривает возможность заказать столик. Что будет дальше? Кто-то уже работает над этим прямо сейчас.

Данные, генерируемые интернетом, уже значительно изменили сферу аналитики, и они будут продолжать это делать. В главе 2 рассказывалось, как стремительно развивается область анализа данных о веб-активности пользователей, а также о преимуществах такого анализа. Анализ комментариев в социальных медиа — это еще один новый аналитический подход, который стал возможным благодаря интернету, поскольку, кроме того, что интернет предоставил потребителям множество преимуществ, порождаемые им данные оказывают огромное влияние на то, как производится анализ и создаются аналитические процессы. Сто лет назад, когда появились телефонные линии, никто не мог подумать об анализе текста комментариев на сайтах социальных медиа, однако он появился.

Анализ данных социальных сетей

Недавней аналитической инновацией, которая представляет собой волновой эффект, является анализ данных социальных сетей в телекоммуникационных компаниях, который изменил управление взаимоотношениями с клиентами. Об этом шла речь в главе 3. В течение многих лет телекоммуникационные компании собирали подробные данные о каждом звонке клиента. Основная цель их сбора заключалась в выставлении счета. Тем не менее на протяжении многих лет эти данные использовались также для других видов анализа.

В последнее время в связи с увеличением вычислительной мощности и развитием аналитики, встроенной в базу данных, телекоммуникационные компании начали серьезно заниматься исследованием сети контактов каждого клиента. Кто из них находится в центре огромного круга людей, которые созваниваются между собой? Какие клиенты изолированы в очень маленьком круге, ограниченном их ближайшими родственниками? Такой анализ не только интересен с точки зрения социальной динамики — он также используется для повышения эффективности мероприятий, направленных на привлечение и удержание клиентов.

Как это происходит? Было доказано, что, как только ключевой пользователь круга уходит к другому поставщику услуги, риск того, что к этому же поставщику уйдут и другие члены этого круга, резко повышается, поскольку есть вероятность, что они последуют за лидером. Сведения о влиянии клиента позволяют принимать более обоснованные решения о том, насколько агрессивно компания должна подходить к привлечению, удержанию и вознаграждению конкретного клиента. Раннее обращение к членам круга может помочь избежать возможных потерь клиентов. Клиентам, которые являются частью большого круга, можно предоставить намного больше удерживающих стимулов по сравнению с тем, на что они могут рассчитывать исходя из их индивидуальных расходов. Такой анализ был невозможен до недавнего расширения масштабируемости, которая теперь стала доступной профессиональным аналитикам. Этот анализ представляет собой волновой эффект, порожденный сбором детальных данных о звонках клиентов для целей выставления счетов.

Применение принципа

Возникает вопрос, каким образом организация может начать использовать собственные волновые эффекты? Ответ: как только вы начнете осваивать встроенную в базу данных аналитику, MapReduce и большие данные, чтобы выйти за рамки, вы сможете воспользоваться волновыми эффектами, порожденными первоначальной инновацией — аналитикой, встроенной в базу данных.

Инновации, касающиеся встроенной в базу данных аналитики и модели MapReduce, порождают волновой эффект, связанный с будущими аналитическими инновациями. Что из того, что было недостижимо прежде, можно сделать теперь благодаря дополнительной мощности и масштабу, которые предоставляют эти подходы? Какая

аналитика стала доступной благодаря возможности освоить новый источник больших данных? Организациям следует раздвинуть свои рамки, перестать фокусироваться на улучшении скорости и начать думать о новых способах анализа, которые доступны сегодня и были просто невозможны раньше.

Воспользуйтесь собственными волновыми эффектами

По мере создания в организации культуры инноваций и открытий темпы внедрения инноваций будут расти. Дополнительным преимуществом станет развивающаяся способность выявлять волновые эффекты, которыми можно воспользоваться, основываясь на предыдущих инновациях. Источник данных, проанализированный сегодня, или процесс, созданный завтра, вполне может привести к созданию чего-то гораздо большего в будущем.

Волновые эффекты проявлялись в различных отраслях по мере развития источников данных, аналитических методов и масштабируемости. Розничные торговцы начали использовать программы лояльности для поощрения клиентов, и полученные данные привели к тому, что розничный бизнес превратился из продуктоориентированного в клиентоориентированный. Данные и аналитика, заложенные в основе кредитной оценки, полностью изменили финансовую сферу. Точность кредитных моделей и наличие полезных данных привели к тому, что для каждого типа потребителя теперь существует целый спектр финансовых продуктов и услуг. Анализ текста переписки с отделом обслуживания клиентов только начинает оказывать влияние на самые разные отрасли. Необходимые данные, инструменты для их анализа и масштабируемые системы стали доступными лишь недавно.

Организация не может добиться успеха, делая то же самое, что и ее основные конкуренты. Она должна опередить их. Следует внедрять инновации не только для обеспечения сиюминутных потребностей, но и для создания волновых эффектов, которые проявят себя в будущем. Если большие данные могут сделать бизнес более эффективным сегодня, представьте себе, насколько более эффективным он будет, когда сработают волновые эффекты.

Принцип 3: сосредоточьте все силы на достижении цели

Для того чтобы подтолкнуть организацию выйти за рамки и использовать волновые эффекты, руководство должно сосредоточить все силы на достижении общей цели. В этом заключается третий принцип.

Без этого не добиться успеха — это доказано и для бизнеса в целом, и для углубленной аналитики и больших данных. Сосредоточиться должны вся организация, каждое подразделение и каждая команда. Необходимо выполнить следующие действия:

Шаг 1: организация должна разделять общее видение того, куда она направляется.

Шаг 2: организации должны быть ясны приоритеты, касающиеся реализации этого видения.

Шаг 3: сотрудники организации должны понимать принцип оплаты своего труда.

Определить правильные цели нелегко. Нелегко распространить видение по всей организации. Нелегко заставить организацию принять новые приоритеты и начать работать над ними. Нельзя в одночасье изменить образ мышления и действий людей, и препятствия, которые необходимо преодолеть, не являются тривиальными. Создание культуры инноваций и открытий требует усилий. Однако в итоге отдача при успешной реализации этой цели может быть огромной.

Создание общего видения

Почему так важно, чтобы команда разделяла общее видение и понимала, для чего это необходимо? Рассмотрим пример с двумя рабочими, занятыми заливкой фундамента для нового дома. Каждого из них спросили: «Что ты делаешь?» Первый ответил: «Я наполняю эти формы цементом, чтобы сделать несущую стену». Второй сказал: «Я строю дом».

Почему эта разница имеет значение? В конце концов, они просто создают несущую стену, не так ли? Они оба закончат эту работу к концу дня, так почему мы должны волноваться о том, как они воспринимают ее? Это важно, потому что без общего видения невозможно успешно внедрять инновации. Иначе у вас будет просто множество людей, строящих «стену дня».

В нашем примере человек, который считает, что он просто должен построить стену в конкретном месте в этот день, не обладает видением того, почему эта стена важна и как она соотносится с целым домом. Второй человек, имеющий более полное видение, гораздо лучше справится с неизбежно возникающими проблемами. Он постарается убедиться, что все корректировки, которые необходимо внести в план, не только

обеспечат появление стены в нужном месте и в нужное время, но и то, что эта стена будет поддерживать остальную конструкцию дома. Нелишним будет убедиться в том, что люди не только понимают свои текущие задачи, но и то, как эти задачи вписываются в общее видение.

Этот же принцип применяется и в сфере аналитики. Очень часто аналитику поручают провести анализ, не давая особых указаний и не обозначая перспективу. Он обрабатывает огромное количество данных и предоставляет подробный анализ с разумными и достоверными результатами. К сожалению, человеку, который поручил проведение анализа, требуется совершенно не то, что было предоставлено. Дело в том, что часто аналитику дают конкретную задачу, а не видение желаемых результатов. Он может сделать лишь то, о чем его попросили, поскольку ему не сказали, что и почему было действительно необходимо. Известный автор и оратор Стивен Кови написал книгу «7 навыков высокоэффективных людей»* (7 Habits of Highly Effective People). Один из них гласит: «Начинай действовать, имея в виду конечный результат»²³.

Аналитику необходимо предоставить видение

Если организация хочет добиться успеха в укрощении больших данных, то работающим в ней аналитикам понадобится видение того, куда они направляются. В противном случае вы получите множество интересных исследований, которые не смогут способствовать развитию бизнеса. Аналитикам стоит давать не задачу на неделю, а долгосрочную перспективу, с учетом которой они будут работать в дальнейшем.

Первый шаг в укрощении больших данных — не просто поручить аналитикам выяснить, что в них заключается. Некоторая часть неуправляемого исследования может быть предусмотрена планом, но это не должно быть единственной его частью. Важно обсудить, что собой представляют эти данные, а также то, как они могут помочь бизнесу. В каких конкретно сферах новые данные могут быть полезны? Какие решения могли бы изменить? Понимая это, аналитики смогут двигаться в правильном направлении.

Аналитику, например, нельзя отдать полную свободу действий над телеметрическими данными видеоигр (см. главу 3). Сначала следует обсудить типы микротранзакций, которые имеют значение в этих играх;

* Кови С. 7 навыков высокоэффективных людей. М. : Альпина Паблишер, 2013. Прим. ред.

далее предоставить любые доступные сведения о предпочтениях игроков; сделать обзор того, что компания хотела бы улучшить или начать делать по-другому. Это дополнительное видение позволит аналитикам добиться положительных результатов гораздо быстрее.

Кем считают себя ваши сотрудники? Теми, кто строит стену здания в определенном месте, или теми, кто строит само здание? Принято ли в вашей организационной культуре задавать вопросы и пытаться понять более общую картину или от людей требуется делать только то, что им сказали? Кто из вашей команды относится к строителям стен, а кто — к строителям дома? Что необходимо изменить, чтобы команда начала делать то, что нужно?

Определение приоритетов

Фактором, без которого организация не может быть успешной, является четкое понимание приоритетов. Как мы уже говорили, сначала организация должна иметь видение того, куда она пытается идти. После этого необходимо установить, как она будет достигать своей цели. Выбранные приоритеты могут кардинально изменить необходимые стратегию и тактику. Например, является ли приоритетом намерение стать крупнейшей компанией в отрасли? Или приоритет — солидная база довольных потребителей? Или максимизация маржи прибыли? Или минимизация оттока клиентов?

От этого зависит выбор стратегии, а также целей, касающихся количества клиентов, потребительской ценности, общей выручки и прибыли. Это, конечно, значительно повлияет на фокус и объем проводимого анализа. Учитывая усилия и время, необходимые для создания культуры инноваций и открытий, организация должна убедиться, что все сотрудники сосредоточены на одной цели и с самого начала имеют одни и те же приоритеты.

Друг одного коллеги — владелец ресторана, в котором состоялся очень успешный «счастливый час». «Счастливый час» позволил людям прийти и купить напитки по низкой цене, однако эта акция не способствовала созданию каких-либо отношений с рестораном. Что произойдет, если дальше по улице откроется бар с еще более дешевыми напитками и более новым интерьером? Хозяин стал размышлять над своими приоритетами.

Он провел эксперимент: предложил оплатить еду и напитки для вечеринки по случаю дня рождения любого человека и его десятирех

гостей при определенных условиях. Именинник должен был присоединиться к фан-клубу ресторана на Facebook и сделать так, чтобы к этому же фан-клубу присоединилось определенное количество его друзей. Кроме того, именинник должен был предоставить имена и адреса электронной почты определенного количества друзей, которых ресторан еще не зарегистрировал.

Идея имела огромный успех. Люди, которые присутствовали на вечеринке, тоже захотели устроить свой праздник в этом ресторане. Между тем база контактов ресторана увеличивалась в геометрической прогрессии. В конце концов владелец прекратил устраивать «счастливые часы» и сосредоточился на днях рождения. Он решил, что для него важнее постоянные клиенты, которые приходят в его ресторан из-за его уникального предложения. Он уже не хотел просто собирать большую толпу людей, которые ищут дешевые напитки.

Этот пример показывает, что подобная история не была бы возможной без электронной почты и сети Facebook, однако это волновой эффект данных инноваций. Владелец вышел за свои рамки, сосредоточился на новой цели и достиг ее. Большая часть его конкурентов по-прежнему сосредоточена на достижении плана продаж напитков. Он сосредоточил свое внимание на создании базы лояльных клиентов, которая будет расти из года в год с каждым их новым днем рождения. Изменение приоритетов заставило его пойти совсем в другом направлении, но зато привело к успеху.

Будьте лучшим... в чем?

Многие организации хотят быть лучшими в своей области. Однако что подразумевается под словом «лучший»? Его можно толковать по-разному, и очень важно, какое именно значение выбрано. После этого необходимо установить приоритеты, соответствующие этому определению. В противном случае усилия окажутся разобщенными и достичь успеха будет трудно. Необходимые виды анализа зависят от целей и показателей, соответствующих понятию «лучший».

Каждой организации необходимо пересмотреть свои приоритеты. Предположим, что организация решила производить анализ текста переписки с отделом обслуживания клиентов и комментариев в социальных медиа. Каковы наиболее важные проблемы, которые необходимо решить? Является ли приоритетом понимание общего настроения комментариев? Является ли приоритетом выявление конкретных людей, которые нуждаются во внимании? Является ли приоритетом

выявление тенденций в плане частоты обсуждения продуктов? Существует так много различных способов использования анализа текста, что без четких приоритетов организации будет трудно добиться успеха.

Привязка зарплаты к видению и приоритетам

Критически важно, чтобы сотрудники знали, от чего зависит их зарплата. Хорошо это или плохо, но зарплата влияет на то, что они делают. Организация должна обеспечить оплату труда и продвижение аналитиков за внедрение аналитических инноваций, в противном случае этого не произойдет. В то же время следует удостовериться, что план компенсаций способствует проявлению нужного поведения. Очень легко ввести изменения, которые могут привести к неожиданным и непреднамеренным результатам.

Некоторое время назад крупная сеть продуктовых магазинов совместно с крупным производителем безалкогольных напитков проводила огромную распродажу. Менеджерам магазина сказали, что они получают бонус в зависимости от объема продаж безалкогольных напитков в их магазинах во время проведения акции. Один магазин показал неслыханно большой объем продаж, значительно превышающий показатели других магазинов сети. Кто-то решил разобраться, как это могло быть.

Что же выяснилось? Менеджер магазина с наибольшими показателями продаж оказался другом менеджеров некоторых местных конкурирующих магазинов. Как ни странно, распродажная цена была ниже, чем цена, по которой конкурирующие менеджеры могли приобрести продукты для своих магазинов. Оказалось, что этот менеджер на самом деле продавал конкурентам палеты с напитками прямо с погрузочной платформы! Конечно же, такие продажи приводили к убытку для его работодателя и к прибыли для конкурентов. Звучит дико, не правда ли?

Люди будут делать то, к чему вы их побуждаете

Если стимулы плохо продуманы, не стоит удивляться, что сотрудники совершают неожиданные поступки, чтобы получить обещанное вознаграждение. Даже самые лучшие из них способны на непредвиденные действия ради награды. Убедитесь, что ваши стимулы способствуют приложению правильного вида анализа к правильным проблемам, а не просто какого-нибудь нового вида анализа, который пришел кому-нибудь в голову.

Был ли менеджер магазина в предыдущем примере нечестен? Трудно сказать. Он, может быть, и нечестен, однако дело не в этом. Менеджеру сказали, что его задачей было продать как можно больше безалкогольных напитков. Если он воспринял эту цель буквально и не имел общего видения, то он оказался не нечестным, а на самом деле весьма изобретательным. Он нашел инновационный способ обеспечить наибольший объем продаж.

Можно утверждать, что на самом деле этот менеджер относится именно к тому типу сотрудников, какие требуются организациям. Он сосредоточился на своей цели, проявил творческий подход в ее достижении и опередил остальных. В данном случае руководство случайно создало сценарий, в котором у данного сотрудника был стимул сделать то, что в действительности не имело смысла. Если энергию и творчество этого сотрудника перенаправить в нужное русло, то он, вероятно, сможет обеспечить положительные результаты.

В мире аналитики обещание бонуса в зависимости от количества созданных моделей может привести к появлению большого количества довольно бесполезных моделей. Лучше предложить бонус, который будет зависеть от эффектов, обеспечиваемых этими моделями. Чем лучше модели, тем больше бонус. Качество должно иметь приоритет над количеством. Если вы предлагаете бонус за обнаружение возможностей использования источника больших данных, то области для исследования должны быть гибкими. Возможно, многие оригинальные идеи не сработали, но другие обнаруженные возможности оказались подходящими. Вознаграждение должно зависеть от нахождения ценности, а не от нахождения ценности для конкретной задачи. В противном случае аналитики рискуют потратить слишком много времени на решение малозначимой задачи только для того, чтобы получить свой бонус.

Применение принципа

Вашей организации необходимо изучить существующее видение, приоритеты и стимулы, а также постоянно пересматривать их и проверять на соответствие друг другу. Нет смысла ни с того ни с сего менять направление. Однако придерживаться неподходящего видения, приоритетов или плана компенсаций может быть столь же неправильным, как и отказ от их первоначальной разработки. Спуск на машине с обрыва —

это не очень хорошая идея, даже если вся команда едина в своем намерении это сделать!

Вернемся к примеру со встроенной в базу данных аналитикой и моделью MapReduce и предположим, что организация решила использовать их для улучшения своих аналитических процессов и освоения больших данных. С чего начать? Нужно поставить команде аналитиков задачу внедрения аналитики в базу данных и реализации среды MapReduce к определенному сроку. Необходимо разработать видение того, как следует изменить аналитические процессы организации в целях обработки больших данных. Определить ясные приоритеты, затем привязать премии и цели к реализации видения, тогда команда будет придерживаться плана.

Ваша организация должна потребовать от аналитической команды выйти за рамки и создать инновационную, беспрецедентную аналитику, которая отличалась бы от используемой в данный момент. Тогда у вас будут результаты. Однако если это останется простым предложением или пожеланием, то ничего не изменится. Для достижения успеха организация должна сосредоточить все силы на единой цели.

Обзор главы

Самые важные уроки этой главы.

- Не прилагая усилий, ваша организация не сможет освоить большие данные. Для того чтобы выиграть, нужно играть! Возьмите на себя обязательство применить к большим данным новые аналитические подходы.
- Широко применяются три принципа, которые также относятся к области углубленной аналитики и больших данных. Они предполагают: 1) выход за рамки; 2) использование волнового эффекта и 3) сосредоточение всех сил на единой цели.
- Ограничивающие рамки не являются исключительно отрицательным явлением. Однако следует постоянно пересматривать их, чтобы проверить актуальность существующих ограничений. Не следует ограничивать себя без необходимости.
- Успех в области аналитики определяется не инструментами и технологиями. Главным фактором успеха являются люди, которые применяют эти инструменты и технологии.

- ▶ Большие данные представляют собой настолько новое явление, что их будущие волновые эффекты еще далеко не ясны. Не упустите шанс воспользоваться преимуществами пока неизвестных волновых эффектов и начните анализировать большие данные уже сегодня.
- ▶ Не концентрируйте свое внимание лишь на улучшении скорости и ищите новые аналитические подходы, которые были невозможны ранее, но которые можно применять сейчас.
- ▶ Аналитикам следует не давать задание на неделю, а поделиться с ними видением того, в каком направлении им нужно работать. Это поможет им оставаться сосредоточенными на цели.
- ▶ Выбор приоритетов способен кардинально изменить стратегию и тактику, используемые для реализации видения. Перед тем как проводить анализ, убедитесь в правильности приоритетов.
- ▶ Разработайте стимулы, которые приводят к получению нужных результатов. Лучше предложить бонус исходя из эффектов разработанных моделей, а не из их количества.
- ▶ Создайте культуру инноваций и открытий. Кто-то в вашей организации должен взять на себя инициативу по внедрению аналитических инноваций и укрощению больших данных. Почему бы не вы?

Заключение

Мыслите масштабнее!

Вы сделали это! Вы узнали о том, что собой представляют большие данные. Вы прочитали об инструментах, процессах и методах, необходимых для их укрощения. Ознакомились с критериями хорошего анализа. Прочитали о том, какие люди и команды нужны для выполнения хорошего анализа. Вы узнали, как внедрить нововведения с помощью центра аналитических инноваций и создания культуры инноваций и открытий.

В ходе нашего разговора мы охватили ряд концепций и методов их применения в целях улучшения аналитики и бизнеса вашей организации. Надеемся, что независимо от уровня подготовки вы получили ряд новых идей, которые можете применить в вашей организации.

Далее перечислены ключевые моменты, на которых следует еще раз сделать акцент, а также описаны меры, которые вы можете принять:

- ▶ Большие данные реальны, и они никуда не исчезнут. Не игнорируйте и не бойтесь их. Пересмотрите с их учетом свои стратегии, касающиеся корпоративных данных и аналитики. Используйте их для обеспечения конкурентного преимущества.
- ▶ Масштабируемость важна как никогда. Убедитесь, что ваша организация использует самые современные технологии, в том числе аналитику, встроенную в базу данных, модель MapReduce и облачные вычисления.

- Необходимы новые процессы. Начните использовать аналитические песочницы, аналитические наборы данных предприятия и встроенный скоринг для обеспечения работы более быстрых и более масштабируемых аналитических процессов.
- Используйте такие новые методы анализа, как анализ текста, групповые и экспресс-модели. Не применяйте старые традиционные методы к новым источникам больших данных.
- Для укрощения больших данных вам потребуются талантливые профессионалы в области аналитики. Выдающиеся специалисты вне зависимости от того, как они себя называют: аналитиками или учеными в области науки о данных, — имеют такие черты, как обязательность, творческий подход, деловая смекалка, навыки презентации и интуиция. Найдите и наймите несколько таких людей.
- Аналитические команды могут быть организованы по-разному. Однако внимание всегда должно быть сосредоточено на предоставлении необходимой информации лицам, принимающим решения.
- Создайте центр аналитических инноваций для помощи в решении задач, связанных с большими данными. Создайте культуру инноваций и открытий. Это упростит процесс укрощения больших данных.

Переход к принятию решений на основе данных происходит уже в течение довольно длительного времени. Количество источников данных и разнообразие средств углубленной аналитики, используемых для принятия таких решений, постоянно увеличивается. Большие данные являются просто очередным дополнением, и их не стоит бояться. Организациям необходимо уже сегодня начать их применять. Нет причин откладывать. Несмотря на неизбежные препятствия и некоторое сопротивление изменениям, укротить большие данные можно, если начать прямо сейчас. Будь то текстовые данные, веб-данные или данные, полученные от датчиков, сегодня уже существуют организации, которые активно собирают, анализируют и принимают решения на их основе.

Организации, которые решили стать лидерами, обнаружат новые бизнес-возможности и разработают новые бизнес-процессы еще до того,

как последователи осознают, что произошло. Редко выпадает шанс оказаться первым в совершенно новой сфере данных и их анализа. Не упустите его! Начинайте искать способы, с помощью которых анализ больших данных может изменить то, как ваша организация ведет свои дела. Воспользуйтесь преимуществами. Чего вы ждете?

Благодарности

Многие люди заслуживают признательности за помощь в написании этой книги. Я благодарю своих коллег из Teradata, SAS и Международного института аналитики, которые побудили меня написать эту книгу, а также знакомых авторов, которые помогли мне понять, во что я ввязался.

Я обязан тем, кто просматривал книгу и делал свои замечания по мере ее написания. Чтение сотен страниц черновика совсем не похоже на праздник! Спасибо за ваш вклад.

И, наконец, спасибо всем профессиональным аналитикам, профессионалам в сфере бизнеса и ИТ-специалистам, с которыми я работал на протяжении многих лет. Все вы помогали мне изучать и применять описанные здесь концепции. Если бы я не имел возможности увидеть их в действии в реальных условиях, то не смог бы написать о них.

Билл Фрэнкс

Об авторе

Билл Фрэнкс — директор по аналитике глобальных партнерских программ компании Teradata, предоставляет информацию о тенденциях в области углубленной аналитики и помогает клиентам понять, какую поддержку им могут оказать Teradata и ее партнеры. Билл курирует центр Business Analytic Innovation Center, который совместно спонсируется компаниями Teradata и SAS, и помогает клиентам внедрять аналитические инновации; кроме того, он содействует компании Teradata в определении правильной стратегии и позиционировании в сфере углубленной аналитики.

Билл преподает в Международном институте аналитики, который был основан ведущим экспертом в области аналитики Томом Дэвенпортом; он активный блогер и оратор. Его блог можно найти по адресу: iianalytics.com/category/faculty-blogs/bill-franks/.

Билл всегда стремился помочь перевести сложную аналитику на язык, понятный бизнес-пользователям, и встроить полученные результаты в процессы организации. Его клиентами являются множество компаний в различных отраслях, от крупных корпораций, входящих в список Fortune 100, до небольших некоммерческих организаций.

Билл получил степень бакалавра в области прикладной статистики в Политехническом университете Вирджинии и степень магистра в области прикладной статистики в Университете Северной Каролины.

Предметный указатель

- 360-градусный обзор 56
- Автоматический сбор оплаты за проезд
 - по платным дорогам, RFID 91
- Автострахование
 - использование телематических данных 80
- Анализ 213
 - «снятие сливок» 221
 - MapReduce 140
 - аналитический набор данных (ADS) 163
 - аналитический набор данных предприятия (EADS) 167
 - важность для бизнеса 228
 - визуализация данных 203
 - встроенный скоринг 176
 - выборка против популяции 229
 - грид-вычисления 138
 - инструменты и методы 185
 - критерии хорошего анализа 218
 - массивно-параллельные системы обработки 121
 - облачные вычисления 131
 - песочница 152
 - песочницы 137
 - постановка проблемы 223
 - предположения против подсчета статистики 232
 - проект R для статистических расчетов 198
 - процессы 151
 - сравнение базовой и углубленной аналитики 219
 - сравнение с отчетностью 213
 - статистическая значимость 224, 225
 - характеристики 216
- Анализ данных 115
 - извлечение, преобразование и загрузка данных (ETL) 118
 - корпоративное хранилище данных 119
 - масштабируемость 116
- Анализ текста 191
- Аналитика с погружением 205
- Аналитическая команда 259
 - взаимное обучение 271
 - взаимодействие с менеджером 271
 - в различных отраслях 260
 - гибридные структуры 268
 - дефицит талантов 262
 - децентрализованные/функциональные структуры 265
 - матричный подход 270
 - нерешительность 261
 - поддержание уровня компетентности 269
 - структуры 264
 - углубленная аналитика 272
 - централизованные структуры 266
- Аналитическая песочница 152
- Аналитические инструменты 195
 - аналитика с погружением 205
 - графические пользовательские интерфейсы (GUI) 196
 - открытое программное обеспечение 199

- точечные решения 197
- Аналитические методы 186
 - анализ текста 191
 - групповые модели 186
 - ранжирование страниц 194
 - совместная фильтрация 194
 - экспресс-модели 188
- Аналитический набор данных (ADS) 163
 - для разработки 164
 - предприятия 167
 - производственный 164
 - традиционный 165
- Аналитический набор данных предприятия (EADS) 167
 - данные в 170
 - использование 175
 - обновление 172
 - особенности 169
 - создание 169
 - сравнение логической и физической структур 171
 - сравнение таблиц и представлений 173
- Базовая аналитика
 - сравнение с углубленной аналитикой 219
- Бизнес
 - важность анализа данных для 228
 - понимание аналитиком 245
- Большие данные 29
 - влияние 29
 - использование 32
 - исследование 42
 - необходимость в анализе 38
 - объединение с традиционными данными 47
 - объем, скорость передачи, сложность 31
 - определение 30
 - отличие от традиционных данных 32
 - процесс извлечения, преобразования и загрузки данных (ETL) 46
 - регулирование 37
 - риски 36
 - стандарты 48
 - структура 39
 - сходство с традиционными данными 34
 - характеристики 50
 - хранилища данных предприятия 47
 - ценность 34
 - эволюция 49
 - эффективная фильтрация 46
- Важность для бизнеса 228
- Веб-данные
 - исследовательское поведение 66
 - конфиденциальность и 61
 - моделирование отклика 72
 - моделирование потерь 71
 - обзор 56
 - обратная связь от потребителей 68
 - оставление корзины 75
 - поведение потребителей 59
 - покупательское поведение 63
 - применение 68
 - пути к покупке и предпочтения 64
 - сегментация клиентов 74
 - следующее наилучшее предложение 69
- Видение
 - зарплата и 320
 - приоритеты и 318
- Видеоигры, телеметрические данные 104
- Визуализация
 - аналитика с погружением 205
 - важность 206
 - данных 203
 - средства 204
- Внешняя песочница 157
- Внутренняя песочница 155
- Встроенные процессы 128
 - массивно-параллельные системы обработки 128
 - язык структурированных запросов (SQL) 128
- Встроенный скоринг 176
 - входные аналитические наборы данных 179
 - выходные данные модели 182
 - информация о модели 180
 - интеграция 177

- пакетное обновление 176
- проверка модели и отчетность 181
- скоринг в режиме реального времени 177
- управление моделями и оценками 179
- язык разметки для прогнозного моделирования (PMML) 178
- Выборка против популяции 229
- Гибридная песочница 159
- Гибридные структуры 268
- Графические пользовательские интерфейсы (GUI) 196
- Грид-вычисления 138
- Групповые модели 186
- Давность, частота и денежная ценность (RFM) 57
- Данные о времени и местоположении 87
 - использование 88
 - маркетинг и 89
 - система глобального позиционирования (GPS) 87
- Данные, полученные от датчиков 101
 - использование 101
- Данные социальных медиа 33
 - пользователи 33
- Данные социальных сетей
 - использование 107
 - определение общей потребительской ценности 108
 - сложность проведения анализа 107
 - телекоммуникации 105
- Децентрализованные/функциональные структуры 265
- Диверсификация, аналитические инновации 289
- Единицы измерения объема данных 117
- Извлечение, преобразование и загрузка данных (ETL) 46, 118
- Инновации
 - в области аналитики 283
 - волновой эффект 311
 - выход за рамки 307
 - диверсификация и аналитические 289
 - изменение точки зрения 289
 - итеративные подходы к внедрению 288
 - ключевые принципы 306
 - необходимость для бизнеса 284
 - общее видение 316
 - определение 287
 - подготовка почвы 304
 - приоритеты и 318
 - риск и 286
 - создание культуры 303
 - сосредоточение всех сил 315
 - традиционные подходы препятствуют внедрению 285
 - центр аналитических инноваций 291
- Интеллектуальные сети 47
 - использование данных 96
- Интернет-журналы
 - полуструктурированные данные 40
- Интернет-транзакции 33
- Интуиция профессионала в области аналитики 253
- Исследовательское поведение 66
- История просмотра веб-страниц
 - конфиденциальность и 38
- Источники больших данных
 - веб-данные 55
 - данные интеллектуальных сетей 95
 - данные о времени и местоположении 87, 88
 - данные отслеживания фишек казино 98
 - данные, полученные от датчиков 95
 - данные радиочастотной идентификации (RFID) 91
 - данные социальных сетей 106
 - текстовые данные 83
 - телематика 80
 - телеметрические данные 104
- Итеративные подходы к внедрению аналитических инноваций 288
- ИТ-специалисты
 - профессионалы в области аналитики 277
- Коммунальные предприятия
 - использование данных интеллектуальных сетей 96

Коммуникационные навыки

- важность представления результатов 251
- навыки презентации и 250
- профессионала в области аналитики 250
- результаты и успех анализа 251
- реклама и 252

Конфиденциальность данных 37, 61

- большие данные и 37

Корпоративное хранилище данных 119

Критерии хорошего анализа 218

Массивно-параллельные системы обработки

- 121
- встроенные процессы 128
- масштабируемость 121
- подготовка данных и скоринг 124
- системы баз данных 123
- функции, определенные пользователем 127
- центральный процессор (CPU) 122
- язык разметки для прогнозного моделирования (PMML) 129
- язык структурированных запросов (SQL) 124

Масштабируемость 116

- MapReduce 140
- грид-вычисления 138
- единицы измерения объема данных 117
- извлечение, преобразование и загрузка данных (ETL) 118
- история 116
- комбинирование аналитических технологий 147
- корпоративное хранилище данных 119
- массивно-параллельные системы обработки 121
- облачные вычисления 131
- система управления реляционными базами данных (СУРБД) 119
- слияние аналитической среды со средой данных 118
- централизация данных 118
- язык структурированных запросов (SQL) 124

Модели

- выходные данные 182

групповые 186

- проверка и отчетность 181
- управление встроенным скорингом 179
- экспресс 188

Моделирование отклика 72

Моделирование потерь 71

Модель проталкивания SQL 127

Национальный институт стандартов и технологий (NIST) 132

Неструктурированные данные 39, 141

MapReduce 141

- анализ текста 141 191
- источники 40

Облачные вычисления 131

- критерии среды 131
- масштабируемость 131
- публичные облака 133
- сравнение с песочницами 137
- частные облака 136

Обратная связь от потребителей 68

Оставление корзины 75

Открытое программное обеспечение 199

Отслеживание имущества, RFID-метки для 92

Отчетность

- сравнение с анализом 213
- характеристики 214

Оценка эффективности рекламы 75

Пассивные RFID-метки 91

Песочница

- анализ данных с помощью 137
- аналитическая 152, 153
- внешняя 157
- внутренняя 155
- выявление новых источников данных с помощью 160
- гибридная 159
- преимущества 154
- сравнение с облачными вычислениями 137, 152

Планирование мощностей 161

Поведение потребителей 33, 55, 57, 59, 60

- исследовательское поведение 66

- конфиденциальность и 61
- обратная связь 68
- покупательское поведение 63
- пути к покупке и предпочтения 64
- Подготовка данных и скоринг 124
 - встроенные процессы 128
 - массивно-параллельные системы обработки 124
 - функции, определенные пользователем 127
 - язык разметки для прогнозного моделирования (PMML) 129
 - язык структурированных запросов (SQL) 124
- Покупательское поведение 63
- Полуструктурированные данные 39, 40
- Пользовательские интерфейсы 196
- Постановка проблемы 223
- Поток кликов 50, 56
- Предположения против подсчета статистики 232
- Презентации, навыки 250
- Производственный аналитический набор данных (ADS) 164
- Промышленные двигатели и оборудование
 - использование данных, полученных от датчиков 101
- Профессионалы в области аналитики 35, 235, 236
 - взаимное обучение 271
 - видение 316
 - деловая смекалка 245
 - интуиция 253
 - использование списка требований 240
 - ИТ-специалисты и 277
 - как ученый в области науки о данных 236
 - команда 259
 - культурная осведомленность 248
 - навыки коммуникации 250
 - навыки презентации 250
 - недостаток 262
 - образование 238
 - отраслевой опыт 239
 - распространенные заблуждения о 237
 - роль 35
 - сертификация 256
 - творческий подход 242
 - уровень детализации при принятии решений 246
 - чистые данные 243
- Процесс извлечения, преобразования и загрузки данных (ETL) 46, 118
- Публичные облака 133
- Пути к покупке и предпочтения потребителей 64
- Радиочастотная идентификация (RFID) 44, 91
 - использование данных 92
 - метки 44
 - метки для автоматической оплаты проезда по платным дорогам 91
 - объединение данных с другими данными 93
 - отслеживание имущества 92
 - отслеживание фишек казино 98
 - пассивные метки 91
 - серийный номер 91
 - уменьшение случаев мошенничества 94
- Ранжирование страниц 194
- Риск и аналитические инновации 286
- Сегментация клиентов 74
- Система управления реляционными базами данных (СУРБД) 119
- Скоринг в режиме реального времени 177
- Следующее наилучшее предложение 69
- Совместная фильтрация 194
- Статистическая значимость 224, 225
- Структуры команд 264
 - гибридные 268
 - децентрализованные/функциональные 265
 - централизованные 266
- Текстовые данные 83, 141
 - анализ 191
 - изменение смысла путем смещения акцента 193
 - инструменты для анализа 84
 - интерпретация 85
 - использование 85

- неструктурированные данные 141
- Телекоммуникации и данные социальных сетей 106
- Телематические данные 80
 - использование 82
 - сбор для целей автострахования 80
- Телеметрические данные
 - видеоигры 104
 - использование 104
- Точечные решения 197
- Традиционные данные
 - объединение с большими данными 47
 - отличие от больших данных 32
 - структура 40
 - сходство с большими данными 34
- Углубленная аналитика
 - обязанности аналитической команды 272
 - сравнение с базовой аналитикой 219
- Управление рабочей нагрузкой 161
- Ученый в области науки о данных 236
- Фишки казино, отслеживание 98
- Функции, определенные пользователем 127
- Хранилища данных предприятия 47
- Централизованные структуры 266
- Центральный процессор (CPU) 122
- Центр аналитических инноваций 291
 - команда 294
 - продукты и услуги сторонних поставщиков 293
 - работа с неудачами 299
 - руководящие принципы 295
 - совет по инновациям 294
 - спонсорство 293
 - сфера деятельности 296
 - технологическая платформа 292
- Частные облака 136
- Чистые данные 243
- Экспресс-модели 188
- Язык разметки для прогнозного моделирования (PMMML) 129
 - встроенный скоринг и 178
 - массивно-параллельные системы обработки 129
- Язык структурированных запросов (SQL) 124
 - встроенные процессы 128, 176
 - массивно-параллельные системы обработки 124
 - функции, определенные пользователем
 - модель проталкивания 127
 - функции, определенные пользователем 127
- MapReduce 140
 - двухшаговый процесс 140, 142
 - масштабируемость и анализ 140
 - обработка неструктурированного текста 141
 - сильные и слабые стороны модели 144
 - фреймворк для параллельного программирования 140
- RFM (давность, частота и денежная ценность) 57
- R, проект для статистических расчетов 200

Примечания

- ¹ Адриан М. Большие данные (Big Data) [Электронный ресурс] // Teradata, 1:11. URL: www.teradatamagazine.com/v11n01/Features/Big-Data/. *Здесь и далее прим. авт.*
- ² Большие данные: следующий рубеж инноваций, конкуренции и эффективности (Big Data: The Next Frontier for Innovation, Competition, and Productivity) // McKinsey Global Institute, май 2011 года.
- ³ Там же.
- ⁴ «Большие данные» — большие возможности (CEO Advisory: “Big Data” Equals Big Opportunity) // Gartner, 31 марта 2011 года.
- ⁵ Эта глава основана на содержании речи, написанной совместно с моей коллегой Ребеккой Букнис. Мы создали документ на данную тему под названием «Повышение качества анализа путем интеграции данных о потоке кликов» (Taking Your Analytics Up a Notch by Integrating Clickstream Data) для международного форума SAS Global Forum 2011.
- ⁶ Комментарий основан на информации, взятой с сайта whatsabyte.com.
- ⁷ Более подробную информацию вы можете получить на сайте www.DMG.org.
- ⁸ Разрешение недоразумений, связанных с облачными вычислениями (Clearing the Air on Cloud Computing) // McKinsey and Company, март 2009 года.
- ⁹ Проект «NIST Рабочее определение облачных вычислений» 8-21-09, версия 15 (NIST Working Definition of Cloud Computing 8-21-09, version 15) // Национальный институт стандартов и технологий.
- ¹⁰ Национальный институт стандартов и технологий. URL: www.nist.gov/itl/cloud/index.cfm.
- ¹¹ Сени Дж., Элдер Дж. Групповые методы в интеллектуальном анализе данных: Повышение точности путем объединения прогнозов (Giovanni Seni and John Elder. Ensemble Methods in Data Mining: Improving Accuracy through Combining Predictions. — Morgan and Claypool Publishers, 2010).
- ¹² James Surowiecki. The Wisdom of Crowds. — Anchor Books, 2005. Издано на русском языке: *Шуровьески Дж.* Мудрость толпы. Почему вместе мы умнее, чем поодиночке, и как

коллективный разум влияет на бизнес, экономику, общество и государство. — М. : Манн, Иванов и Фербер, 2013

- ¹³ The Comprehensive R Network (CRAN). URL: http://cran.r-project.org/doc/FAQ/R-FAQ.html#Why-is-R-named-R_003f/.
- ¹⁴ Тафти Э. Р. URL: www.edwardtufte.com/tufte/books_vdqi/.
- ¹⁵ Тафти Э. Р. Тафти. URL: www.edwardtufte.com/tufte/minard/.
- ¹⁶ Immersive Intelligence Colleagues. URL: <http://im-tel.org/>.
- ¹⁷ Дэвенпорт Т., Харрис Д. Организация аналитиков (Organizing Analysts) // Международный институт аналитики, резюме вебинара, 23 июня 2009 г.
- ¹⁸ Мой блог можно найти по адресу: iianalytics.com/category/faculty-blogs/bill-franks/.
- ¹⁹ Innovation // Словарь Merriam—Webster. URL: www.merriam-webster.com/dictionary/innovation.
- ²⁰ Онлайн-магазин Think Geek. URL: www.thinkgeek.com/geektoys/cubegoodies/c208/.
- ²¹ Компания SAS. URL: www.sas.com/company/about/history.html.
- ²² Copernicus, Nicolaus (Коперник, Николай) // New World Encyclopedia. URL: www.newworldencyclopedia.org/entry/Nicolaus_Copernicus.
- ²³ Стивен Кови. URL: www.stephencovey.com/7habits/7habits-habit2.php.

Максимально полезные книги от издательства «Манн, Иванов и Фербер»

Заходите в гости: <http://www.mann-ivanov-ferber.ru/>

Наш блог: <http://blog.mann-ivanov-ferber.ru/>

Мы в Facebook: <http://www.facebook.com/mifbooks>

Мы ВКонтакте: <http://vk.com/mifbooks>

Предложите нам книгу: <http://www.mann-ivanov-ferber.ru/about/predlojite-nam-knigu/>

Ищем правильных коллег: <http://www.mann-ivanov-ferber.ru/about/job/>

Фрэнкс Билл

Укрощение больших данных

Как извлекать знания из массивов информации
с помощью глубокой аналитики

Издано при поддержке компании Teradata

Главный редактор *Артем Степанов*

Ответственный редактор *Мария Красовская*

Арт-директор *Алексей Богомолов*

Редактор *Татьяна Собко*

Дизайн переплета *Станислав Леонтьев*

Верстка *Вячеслав Лукьяненко*

Корректоры *Лев Зелексон, Юлия Молокова*